Optimization of Genomic Prediction of Single-Cross Performance in Maize

A DISSERTATION SUBMITTED TO THE FACULTY OF

UNIVERSITY OF MINNESOTA

BY

Dnyaneshwar Chandrakant Kadam

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS

FOR THE DEGREE OF

DOCTOR OF PHILOSOPHY

Aaron J Lorenz, Advisor

September, 2017

Acknowledgements

I am foremost thankful to my advisor Dr. Aaron Lorenz for providing me the opportunity to work on this project and all his support and supervision throughout the course of this project. I am also thankful to my advisory committee at University of Minnesota, Dr. Candice Hirsch, Dr. Rex Bernardo and Dr. Shengqiang Zhong, and University of Nebraska, Dr. Ismail Dweikat, Dr. Stephen Kachman, and Dr. Thomas Hoegemeyer for their helpful suggestions and comments.

I am grateful to Indian Council of Agricultural Research for awarding me the fellowship for Ph.D. study, University of Nebraska and University of Minnesota for graduate education and necessary infrastructure for this project, Dr. Martin Bohn from University of Illinois for providing one dataset used in this research, and Dr. Ronald Haarmann from Pioneer for graciously allowing field experiments of hybrids at their York location for two consecutive years.

I would also like to thank Lorenz Lab members and friends for their help in this project as well as making my journey through graduate school pleasant.

Finally, I would like to express by gratitude to my family and relatives for their constant love and support throughout my education.

Abstract

Prediction of single-cross performance has been a major goal of plant breeders since the beginning of hybrid breeding because it is not feasible to evaluate all single-cross combinations between parental inbreds in a hybrid breeding program. Recently, simulation and experimental studies have shown great promise of genomic prediction of single-cross performance. However, further investigations are needed for optimal implementation of genomic prediction for single-cross performance. The objectives of this dissertation were to (1) examine the potential of genomic prediction of single crosses in the early stages of hybrid breeding pipeline, (2) evaluate the nonparametric models for genomic prediction of early-stage single crosses and (3) optimize the training set composition for genomic prediction of early-stage single crosses. Two different datasets consisting of 481 and 312 single crosses generated between random set of recombinant inbred lines (RILs)/doubled haploid lines (DHLs) derived from series of biparental families belonging to Iowa Stiff Stalk Synthetic (BSSS) and Non-Stiff Stalk Synthetic (NSSS) heterotic group were used. All the parental RILs/DHLs were genotyped using genotyping by sequencing approach. The accuracies of genomic prediction were substantially higher than topcross-based prediction commonly used in the early stages hybrid breeding. Moreover, genomic prediction outperformed phenotype-based prediction when only one or none of the parents of single crosses were tested. The mean genomic predictive abilities for T2, T1F, T1M, and T0 single crosses were 0.67, 0.60, 0.55, 0.46 for GY and 0.84, 0.74, 0.74, 0.63 for PH correspondingly. Genomic best linear unbiased prediction (GBLUP) and three nonparametric models namely reproducing

kernel Hilbert space (RKHS), support vector regression (SVR) and neural network (NN) provided similar predictive abilities. Genetic relationship and training set (TRS) size in addition to the number of tested parents of single crosses considerably influenced the predictive abilities. Expected prediction accuracies based on prediction error variance (PEV) agreed well with empirical prediction accuracies when population structure was accounted. Genomic prediction models constructed on TRS optimized with PEV mean and coefficient of determination (CD) mean criteria provided increased predictive ability than stratified and randomly sampled TRS. Overall, the results of this study suggest that genomic prediction of early-stage single crosses with TRS optimization using PEV and CD mean criteria has great potential to redesign hybrid maize breeding and increase its efficiency.

Table of Contents

List of Tables

List of Figures

Chapter 1: Introduction and Literature Review

INTRODUCTION

Maize is one of the most important cereal crop worldwide (Shiferaw *et al.* 2011). It belongs to grass family Poaceae (Gramineae) which also includes other major cereals such as wheat and rice. Maize is further classified into genus Zea, a group of annual and perennial grasses native to Mexico and Central America. The genus Zea comprises wild taxa, collectively known as teosinte (*Zea* spp.) and domesticated maize (*Zea mays* L. ssp. *mays*). Genetic studies have indicated that maize is domesticated directly from Mexican annual teosinte *Zea mays* ssp. *parviglumis*, native to the Balas river valley in south-eastern Mexico (Matsuoka *et al.* 2002). Maize being a versatile crop is grown over a wide range of agro climatic zones (Hake and Ross-Ibarra 2015). The United States, China, Brazil are the top three maize producing countries in the world. Important uses of maize include food, animal feed and bioenergy production (Ranum *et al.* 2014). Maize is also an exciting model organism for biological research including plant domestication, genome evolution, epigenetics, heterosis and quantitative inheritance (Strable and Scanlon 2009).

Development of hybrid maize is one of the landmark achievement in the history of plant breeding (Duvick 2001). Shull (1908, 1909, 1952) first proposed a method for exploitation of heterosis through the development of single-cross hybrid. His method consist of two steps. In the first step, pure line are developed by self-fertilization and the second step consist of identification superior hybrid combination among the pure lines.

This method is referred as “pure line method of corn breeding”. Although Shull’s method has become a standard procedure hybrid breeding programs, several modifications have occurred over the past hundred years to efficiently generate lines and identify superior hybrid combination between them. The main modifications include organization of inbreds into heterotic groups to increase the probability of obtaining superior hybrids (Reif *et al.* 2005), population improvement methods to increase the frequency of lines having good potential for hybrid performance (Comstock *et al.* 1949), topcross test-based screening of lines for hybrid performance (Jenkins and Brunson 1932), doubled haploid (DH) technology to rapidly generate homozygous lines (Rober *et al.* 2005).

Contemporary hybrid maize development program consist of two overlapping stages 1) line development and 2) hybrid evaluation (Figure 1). Development of lines is most commonly carried out by selfing in a pedigree breeding method or DH technology. The crosses between elite lines in each heterotic group are typically used to derive recombinant inbred lines (RILs) or DH lines (DHLs) (Bauman 1982; Mikel and Dudley 2006). Prior knowledge of performance of parental lines in earlier breeding cycles and pedigree relationships is used to determine the potential of specific cross. Simulation studies indicate that selection of parents of crosses is far more important than number of crosses and number of lines within each cross (Bernardo 2003; Wegenast *et al.* 2008). For RILs development via pedigree method, selection and selfing is initiated in F2 population without any genetic recombination as the effects of random mating in F2 population before selfing are not conclusive (Bernardo 2002). DHLs are typically produced from F1 plants instead of F2 plants to shorten the length of breeding cycle

(Longin *et al.* 2007). Recent studies, however, suggest to produce DHLs from F2 plants preferably after selecting for disease and insect resistance, marker associated traits or topcross performance (Wegenast *et al.* 2008; Bernardo 2009).

Characterization and selection of lines involves sequential testing. Initially lines are selected based on per se performance and topcross test while selections in the advanced stages are performed by general combining ability (GCA) and specific combining ability (SCA) evaluation in hybrid combinations. The selections based on per se performance is carried out for traits having reasonably high heritability and are considered necessary for hybrid performance such as maturity, plant canopy architecture, ear size, grain quality and resistance to certain pests and diseases. In case of RILs development, topcross testing is commonly performed in F3 and F4 generations after the lines having poor per se performance are discarded (Hallauer and Miranda 1988). Similarly, only those DHLs having suitable per se performance are evaluated in topcross test. Typically, two generations of topcross testing are conducted (Heffner *et al.* 2010). With each round of topcross test, the number of lines advanced decreases while number of testers used increases. Narrow-based testers such as elite inbred from opposite heterotic group are commonly used for topcross test. Theoretical and empirical results shows that elite inbred from opposite heterotic group generate testcross genetic variability as large as when poor performing tester is used (Hallauer and Lopez-Perez 1979; Bernardo 2002). Additionally, they are more practical because superior hybrid combination can be commercialized in a short period of time.

Lines selected based on topcross performance are evaluated for GCA and SCA in hybrid combinations. Lines retained after GCA and SCA testing are further evaluated in more hybrid combinations at multiple locations. Resources are allocated to evaluate as many lines as possible at topcross stage with intense selection, while at later stages, emphasis is placed on testing fewer hybrid combinations at many locations. The RILs/DHLs with superior GCA and stability identified from multilocation hybrid evaluation are often recycled as parents to develop source populations for line development (Smith 2004). A typical structure of commercial hybrid maize breeding program based on DHLs is depicted in Bernardo (2002).

LITERATURE REVIEW

Hybrid Prediction Problem

Currently, heterotic groups are well established in maize and single-cross hybrids are exclusively made by crossing RILs/DHLs across heterotic groups (Reif *et al.* 2005). This greatly facilitates the hybrid development. However, as the number of lines to be tested are increasing over time especially with advances in DH technology, their evaluation in all possible hybrid combination is challenging. For example, if the breeder has just 100 RILs/DHLs in each heterotic group, the total number of hybrid combination to evaluate is 10,000. Therefore, evaluation of lines for hybrid performance has been the most expensive and critical phase in hybrid maize breeding. If the promising single crosses could be identified without the need to generate and test several thousand

possible single-cross combinations, the efficiency of hybrid breeding would be greatly enhanced (Schrag *et al.* 2009).

Approaches for Hybrid Prediction

In view of its potential to accelerate the hybrid breeding, prediction of hybrid performance has been the major goal of numerous studies. Below is the brief description of different approaches investigated for hybrid prediction.

Inbred per se performance

The possibility of inbred per se performance as indicative of its performance in hybrid combination is desirable to reduce the number of hybrid combinations to be made and tested. Many correlation studies between inbred and hybrid traits were undertaken in the past. The results of these studies were summarized in Hallauer and Miranda (1988). The correlations between inbred per se performance and hybrid performance were variable depending on the traits considered, environment, and tester used. In general, the correlations were relatively high for simple traits such as morphology, ear traits, maturity, quality characters etc. but were relatively low for complex trait such as grain yield. The poor correlations between inbred and hybrid grain yield were due to presence of strong nonadditive effects (Hallauer 1990) and genotype by environment interaction for this trait (Bernardo 1991). Environmental factors also influence the correlation between inbred and hybrid traits. Betran *et al.* (2003) found stronger correlation between inbred and hybrid traits under severe drought stress than under unstressed environments. Smith (1986) reported low theoretical correlation between inbred per se performance and

testcross performance for traits controlled by large number of genes showing complete dominance. The type of tester influenced this correlation, the correlations were low but greater with good unrelated tester than with a good related tester. It is now generally agreed that effective selection can be made on inbred per se performance for certain traits, but evaluation in hybrid combinations is required to identify the lines with best breeding values for complex traits (Hallauer and Miranda 1988).

General combining ability

Combining ability is defined as the capacity of individual to inherit superior performance to its offspring. Initially, it was a general concept used for classifying the lines relative to its performance in hybrid combinations. Sprague and Tatum (1942) refined the concept of combining ability into GCA and SCA. They defined GCA as the average performance of line in a series of hybrid combinations and, SCA as those instances in which certain hybrid combinations are either better or poorer than would be expected on the basis of average performance of parental lines included. Generally, GCA is considered to be an indicator of additive genetic effects, while SCA is related to the nonadditive genetic effects.

Several techniques have been proposed for the estimation of combining ability (Fasahat *et al.* 2016). The two main techniques include topcross test suggested by Davis (1927) and developed by Jenkins and Brunson (1932) and diallel analysis by Griffing (1956). Evaluation in diallel schemes is ideal as it provides information of both GCA and SCA. However, number of single crosses required for diallel analysis inceases

exponentially with increase in the number of lines. Therefore, diallel cannot be conducted practically with substantial number of lines. When the lines belongs to distinct groups like heterotic groups in maize, combining ability evaluation is performed in two factor factorial design (Comstock and Robinson 1948). Analysis of factorial can also provide information on both GCA and SCA. However, as with diallel, it is difficult to evaluate large number of lines in complete factorial because total number of crosses becomes unmanageable.

Topcrossing with appropriate tester has been a simple and widely used approach to evaluate the combining ability of lines (Jenkins and Brunson 1932). The tester genotypes can be classified into two groups namely broad genetic base tester (e.g. Synthetic variety) and narrow genetic base tester (inbred and single-cross hybrid). A broad genetic base tester is considered for GCA evaluation while narrow genetic base tester is useful for SCA evaluation. The average performance of lines with more than one inbred tester is also considered as the measure of lines GCA.

The sum of parental lines GCA estimated using performance in hybrid combinations or topcross test is a simple and established approach to predict single-cross performance (Cockerham 1967; Melchinger *et al.* 1987). The correlations between parents GCA and single-cross performance ranged from 0.68 - 0.94 in different experimental studies in maize (Schrag *et al.* 2006, 2007). Topcross based screening of lines has limitation that it takes longer time to develop commercial hybrid due to additional years of topcross test. Also, all possible single-cross combinations among

available lines cannot be evaluated due to discarding of lines based on topcross test which could include some potential hybrids.

Best linear unbiased prediction

Best linear unbiased prediction (BLUP) is used in linear mixed model for the estimation of random effects. BLUP was derived by C. R. Henderson for the prediction of breeding values in animal breeding (Henderson 1984). Bernardo (1994, 1995, 1996a, 1996b) showed the usefulness of BLUP with interpopulation genetic models involving both GCA and SCA for prediction of untested single crosses. The BLUP approach for single-cross performance used information on genetic relationships among the parental lines, based on coefficient of coancestry estimated from pedigree or molecular marker data. Later, Bernardo (1998) extended aforementioned BLUP (T-BLUP) approach to make use of both trait and marker data (TM-BLUP). In TM-BLUP approach, covariances associated with quantitative trait loci (QTLs) were modelled by inferring the identity by descent of unobservable QTLs from flanking markers. In experimental study, however, T-BLUP and TM-BLUP resulted in similar prediction accuracies which was explained by presence of large number of QTLs for grain yield (Bernardo 1999b). Results indicated that BLUP is useful for routine prediction of single-cross performance (Bernardo 1999a). Pedigree-based BLUP, however, has limitation that it is ineffective in comparing inbreds developed from single biparental population, as they possess the same pedigree (Bernardo 2002). Also, pedigree information is not always available in the breeding program, and, when available does not always have high reliability. The hybrid

prediction accuracies of marker based BLUP could be further improved by use of genome wide dense marker data available in the recent years (Xu *et al.* 2014).

Genetic distances based on molecular markers

With the discovery of molecular markers, genetic distances (GD) between parental lines based on random DNA markers were tested for predicting hybrid performance. Quantitative genetics theory suggests that the amount of heterosis is a function of the allelic diversity between two parents (Falconer *et al.* 1996). Therefore, GD based on molecular markers seemed to be a logical approach for prediction of hybrid performance. However, correlations between hybrid performance and GD for inter-heterotic group hybrids have been very low and/or inconsistent (Melchinger 1999; Lee *et al.* 2007). Two possible sources of these low prediction accuracies include (1) loose linkage between heterotic QTL and the molecular markers used to estimate GD and (2) opposite linkage phases between the QTL and marker alleles as generally expected with inter-heterotic group hybrids (Charcosset *et al.* 1991; Bernardo 1992). Commercial hybrids in maize consist of only inter-heterotic group single crosses, making them the only ones relevant for prediction in breeding programs.

Hybrid performance associated markers

In a modified approach, molecular markers were first tested for association with hybrid traits. Subsequently, a sum of the effects of significantly associated markers ("total sum of selected markers TCSM") was used for prediction of hybrid performance and SCA (Vuylsteke *et al.* 2000). However, this approach was found to be less effective

than established GCA method (Schrag *et al.* 2006). Further, Schrag *et al.* (2007) extended TCSM approach to account for the multiple testing, missing marker data, multiple alleles, which they referred as "total effects of associated markers" (TEAM). The prediction accuracies were still substantially lower than GCA method. Also, extending the GCA predictions with SCA estimates from associated markers did not improve the prediction accuracy (Schrag *et al.* 2006).

Genomic Selection

Genomic selection (GS) is defined as the selection for a trait of interest using large number of genomewide markers (Meuwissen *et al.* 2001). The main difference between marker assisted selection (MAS) and GS is that only the markers that are significantly associated with QTL are used in MAS, while all the markers are used simultaneously without significance testing in GS. An implicit assumption in GS is that all QTLs are in LD with at least one marker. GS has become feasible in the recent years because of following advances in the science and technology: 1. Efficient methods to genotype large number of SNPs discovered by whole genome sequencing (Thomson 2014), 2. Successful application of statistical methods to handle the high dimensional marker data (Gianola *et al.* 2010; de los Campos *et al.* 2013), and 3. Availability of high capacity computational resources (Wu *et al.* 2011). GS is expected to be more effective than MAS especially for traits controlled by many small effect QTLs because the marker effects are less biased in GS compared to MAS as all the marker effects are estimated simultaneously. Also, the proportion of genetic variance explained by markers is larger in

GS than MAS as small effect QTLs that do not meet the significance threshold are missed in MAS (Jannink *et al.* 2010).

The central process of GS consist of two steps. First step is the construction of genomic prediction equation by using marker and phenotypic data on the subset of individuals called the training set (TRS). Second step consist of using the prediction equation to calculate genomic estimated breeding value (GEBV) for a set of selection candidates having only marker data called test set (TS) and then to select the best candidates based on their GEBV. The main challenge in building a genomic prediction model is that number of molecular markers (i.e. predictors) is typically far more than number of individuals in the TRS (i.e. observations) known as "large *p* and small *n*" problem. In this situation, ordinary least square estimates of parameters have large variances leading to poor predictive ability. To confront this problem, a slew of alternative statistical models have been employed with different underlying assumptions. These models can be broadly separated into two categories: parametric and non-parametric. Prominent features of different parametric and nonparametric GS models as well as software packages used to implement them are provided in Table 1. Briefly, parametric models on priori assume a certain form relationship between genetic value and marker covariates. The marker effects are estimated either using shrinkage or combination of shrinkage and variable selection procedure. The common parametric models include ridge regression best linear unbiased prediction (RRBLUP) and Bayesian models. In RRBLUP, marker effects are assumed to be random and normally distributed with common variance resulting in equal shrinkage of their effects. The RRBLUP model

can be implemented in mathematically equivalent but computationally efficient form GBLUP (Habier *et al.* 2007). GBLUP uses genomic relationship matrix (GRM) derived from marker genotypes among the individuals (VanRaden 2008) instead of calculating individual marker effects. This reduces the number of equations required to be solved from *p* to *n*. In contrast to RRBLUP, Bayesian models fits marker specific variances resulting in unequal shrinkage of their effects. Bayes A assign *t*-distribution for marker effects which causes strong shrinkage towards zero for small estimates of marker effects and less shrinkage for sizable estimates of marker effects. Bayes B also assign *t*-distribution for marker effects but, additionally, can set large number of marker effects to zero. Bayes C method is similar to Bayes B except that it assign normal distribution to nonzero marker effects.

Non-parametric models take a different approach by not making strong assumption about the form of relationship between markers and genetic value. Instead, these models seek the form that best fits in TRS data while maintaining some generality for new data. In another words, their main focus is on prediction. These models are, therefore, expected to capture nonaddtive effects without explicitly modelling them and provide better prediction of phenotypes for complex traits (Gianola *et al.* 2006, 2010). Some commonly used nonparametric GS models include reproducing kernel Hilbert spaces (RKHS), support vector regression (SVR) and neural network (NN). RKHS model uses kernel function to define the genetic relationship between individuals which enables to perform non-linear regression in high dimensional feature space. SVR and NN are

machine learning methods which are used to address large *p* small *n* problem in many field.

The predictive performance of GS models is typically evaluated by cross-validation (CV) technique. CV is applied in a number ways depending on the specific objective of the study. CV design includes 1. *k*-fold validation, 2. Repeated random sampling validation, 3. Across cycles/generations validation, 4. Across populations validation, and 5. Across environments validation. In *k*-fold CV, the entire data set is randomly divided into *k* folds. Out of which, *k*-1 folds are used to train the GS model and remaining fold is used as TS. The procedure is repeated till each fold is included in the TS one time. Repeated random sampling CV involves random sampling of data into TRS (e.g. 90 percent) and TS (e.g. 10 percent) several times. Other CV designs (i.e. 3,4,5) involves stratification across cycles/generations, populations, environments correspondingly.

The accurate assessment of prediction accuracy is an important component in evaluating predictive performance of different models. Ideally, the accuracy of GS is the correlation between true breeding value (TBV) and GEBV. However, in practice, TBV is unknown. Considering cor ($\hat{g}$, y) = cor ($\hat{g}$,g) × cor (g,y), where y is vector of phenotypes; $\hat{g}$ is a GEBV and g is TBV, the prediction accuracy is estimated as cor ($\hat{g}$, y) / *h* because h^2 (heritability) = (cor (g,y))^2 (Legarra *et al.* 2008; Hayes *et al.* 2009).

GS accuracy is affected by several factors acting interconnectedly. These factors mainly include genetic relationship, heritability, TRS size, marker density and statistical

methods. Below we describe the connection between individual factor and the prediction accuracy. A detailed discussion of the effects of different factors on GS accuracy can be found in Lorenz *et al.* (2011) and Lin *et al.* (2014).

Genetic relationship

The genetic relationship between TRS and TS is the most important factor influencing the accuracy of genomic prediction. The TRS needs to representative (i.e. closely related) to the TS in order to get good prediction accuracy. The closer genetic relationship benefits the prediction accuracy in two ways 1. It reduces the effective population size generating strong long range linkage disequilibrium (LD) between maker and QTL, and 2. Estimated marker effects are well predictive in TS due to related genome structure. The prediction accuracy is, therefore expected to be highest for training and prediction within population (full sib relationship), followed by populations connected by one shared parents (Half sib relationship). Empirical GS studies in many crops including maize have stressed the importance of genetic relationship for obtaining good prediction accuracy (Albrecht *et al.* 2011; Riedelsheimer *et al.* 2013; Jacobson *et al.* 2014; Albrecht *et al.* 2014).

Heritability

Heritability is an important determinant of achievable prediction accuracy. High heritability enables to estimate marker effects accurately because phenotypic variation is mostly composed of genetic variation with only little confounding effect of environmental factors. Highly significant correlation has been observed between

heritability and prediction accuracy in empirical studies in maize (Lorenzana and Bernardo 2009; Jacobson *et al.* 2014). Although the accuracy of both GS and phenotypic selection is affected by heritability, GS becomes more efficient over phenotypic selection with decrease in heritability (Bernardo and Yu 2007; Viana *et al.* 2016). The genetic relationship information and LD between markers and QTLs enable GS to outperform the phenotypic selection under low heritability situation.

Training Set size

Increasing the TRS size allows accurate estimation of marker effects and consequently enhance the prediction accuracy. Positive correlation between TRS size and prediction accuracy has been reported from studies in maize (Lorenzana and Bernardo 2009; Albrecht *et al.* 2011; Zhao *et al.* 2012). The empirical studies in maize indicate that minimum TRS size of about 50 -100 when predicting within a biparental population (full sib) (Lorenzana and Bernardo 2009; Albrecht *et al.* 2011; Riedelsheimer *et al.* 2013) and about 300 - 400 when predicting for populations related by at least one common parent (half sib) (Zhao *et al.* 2012) are required to obtain prediction accuracy above 0.5 assuming moderate to high heritability. It is important to note that increasing the genetic relationship between TRS and TS is more effective way to increase the prediction accuracy than increasing the TRS size by adding less related individuals. Alternatively, reasonable TRS size is required to obtain reliable prediction even under close genetic relationship between TRS and TS.

Number of markers

The number of markers required to obtain optimal prediction accuracy depends on LD in the population under consideration. If the LD is high, less markers are required and vice versa. The prediction accuracy benefits from increasing the number of markers until sufficient genome coverage is attained. Also, increasing marker density is beneficial only with corresponding increase in TRS size. In maize, about 100 markers for GS within biparental population and 200 - 400 markers for GS with multiple interconnected populations are suggested to get optimal prediction accuracy (Lorenzana and Bernardo 2009; Zhao *et al.* 2012). Due to readily availability of cheap and abundant genome wide SNPs now a days, marker density would not be limiting factor to obtain maximum achievable prediction accuracy. Type of markers used can also influence the prediction accuracy. Solberg *et al.* (2008) reported that three times higher SNP density is required to obtain prediction accuracies comparable to SSR. This is because SSR have multiple alleles and therefore contain more information. The multi-allelic system of SSR can be mimicked by constructing haplotype containing multiple alleles. The improvement in prediction accuracy using haplotype is however minimal especially when SNP density is high (Calus *et al.* 2008). In another study, Poland *et al.* (2012) found greater GS accuracy using SNPs obtained from GBS than DArT marker.

Statistical model

The prediction accuracy of different GS models depends on genetic architecture of the trait and LD structure in the population. Simulation results indicate that RRBLUP and GBLUP rely strongly on kinship while Bayesian models focus more on LD between

marker and QTL than on kinship (Habier *et al.* 2007; Zhong *et al.* 2009). Thus, if there are only few major effect QTLs for a trait, Bayesian models can provide better accuracy over RRBLUP and GBLUP. Alternatively, if there are many small effect QTLs both methods can achieve similar prediction accuracies. The results of empirical studies, however, showed comparable performance of both types methods across different types of trait architectures (Lorenzana and Bernardo 2009; Moser *et al.* 2009). When strong long range LD exist in a population, the effects of major QTLs can be captured by markers well apart from the QTL (i.e. distribution of QTL effect) resulting in good prediction accuracies of RRBLUP. When the nonadditive gene effects are important for a given trait, simulation and some experimental studies indicate the better performance of nonparametric models over parametric methods (Heslot *et al.* 2012; Pérez-Rodríguez *et al.* 2012; Howard *et al.* 2014; Jiang and Reif 2015).

One of the key issue in implementing the GS in a breeding program is how to design TRS to obtain optimal prediction accuracy with minimum phenotyping. One of the approach would be to use phenotypic and genotypic data on genetically related individuals for model calibration. Typically, several lines are evaluated very year in a breeding program. Hence, the phenotypic and genotypic data from multiple differentially related populations is likely to be available even before a new cross is made. Selective phenotyping is another important alternative to reduce the phenotyping expenses involved in GS. Here, the objective is to select minimum number of individuals that are best suited to build the genomic prediction model. Some of the criteria for selection of individuals include stratified sampling, minimizing the prediction error variance (PEV)

and maximizing the reliability (i.e. coefficient of determination) (Rincent *et al.* 2012; Isidro *et al.* 2015).

Comparison of Phenotypic, Marker Assisted Recurrent Selection and Genomic Selection

The relative efficiencies of phenotypic selection, marker assisted recurrent selection (MARS) and GS were compared in simulation and experimental studies in maize. Bernardo and Yu (2007) first showed the effectiveness of GS in plant breeding. They simulated MARS and GS for testcross performance using DHLs derived from single biparental population. The response to GS was 18 to 43% larger than MARS across different numbers of QTLs and levels of heritability. Later, Massman, *et al.* (2013b) provided the first empirical proof of advantage of GS over MARS in crops. Their experiment involved two cycles of GS and MARS for testcross performance for stover and yield indices in a population consisting of 233 RILs derived from B73 and Mo17. The realized gains were 14 to 50% larger with GS compared to MARS. Further, Beyene *et al.* (2015) compared the genetic gain for grain yield in eight biparetal populations under managed drought stress conditions using GS vs pedigree selection. The response to GS was two to four times higher than pedigree selection. The average gain from GS per cycle across eight populations was 0.086 Mg ha^{-1}. They also reported that hybrids derived from cycle 3 (C3) produced 7.3% higher grain yield than those developed through pedigree breeding. Recently, Vivek *et al.* (2017) reported a study on genetic gain under drought conditions using phenotypic selection and GS in two biparental populations. C1

was formed by intermating the top 10% families selected based on testcross performance. Subsequently, C2 derived based on phenotypic selection (C2-PS) and GS (C2-GS). Topcrosses of C2-GS showed 4 - 43% higher grain yield than those of C2-PS. In another recent study, Zhang *et al.* (2017) first applied rapid cycle GS to multiparental population derived from 10 elite maize parents for four recombination cycles. The realized genetic gain with GS cycles (C1-C4) was 0.225 tonn ha^{-1} $cycle^{-1}$ which is equivalent to 0.100 tonn ha^{-1} $year^{-1}$.

Genomic Selection for Hybrid Performance

Several studies have examined the potential of GS at different stages of hybrid maize breeding including per se performance, topcross performance, and single-cross performance (Table 2). GS for per se performance of lines has limited usefulness as the value of line in hybrid breeding is determined by its performance in hybrid combination. However, GS can be benefical for traits such as disease resistance which are evaluated on line per se. Technow *et al.* (2013) investigated the accuracies of genomic prediction of northern corn leaf blight resistance among inbreds belonging to dent and flint heterotic group. The prediction accuracies were low to moderate. They found considerable benefit of increasing the training set size within heterotic groups as well as by combining inbreds across two heterotic groups. Riedelsheimer *et al.* (2013) evaluated the prospects of combining multiple differently related populations into TRS for predicting per se performance of lines for five traits including Gibberella ear rot severity and three kernel yield component traits. They observed considerable decline in predictive ability when full

sib lines were replaced by half-sib lines, but significant predictive abilities were obtained when half-sib lines were available from both the parents instead of only one parent of validation population. Also, some negative effect of combining unrelated populations into TRS was observed.

The genomic prediction studies for topcross performance have looked at the effect of different factors such as TRS size, marker density, prediction within vs across populations, prediction across testers. Generally, the topcross prediction accuracies were benefited by increase in TRS size and number of markers (Lorenzana and Bernardo 2009; Albrecht *et al.* 2011; Zhao *et al.* 2012). Prediction within a biparental population is an ideal scenario because of close relationship between TRS and TS and long range haplotype blocks which creates high LD between marker and QTL. As expected, the mean topcross prediction accuracies for within biparental population were moderate to high (Lorenzana and Bernardo 2009; Albrecht *et al.* 2011; Zhao *et al.* 2012). However, in this scenario, there is the need to phenotype a subset of individuals from the same population which increases the time and cost before genomic prediction can be performed. Also, individual population sizes need to be sufficiently large to reliably perform within population predictions (Schulz-Streeck *et al.* 2012). It would therefore be advantageous if performance of lines within a biparental population could be predicted before that population is phenotyped. In this context, some studies investigated the effect of estimating marker effects across populations to predict within each population (Albrecht *et al.* 2011; Zhao *et al.* 2012; Jacobson *et al.* 2014). The prediction accuracies were similar or slightly lower than within population prediction when the topcross

information of half-sib lines from both the parents were available. The prediction accuracies were severely decreased when topcross information of half-sib lines from only one or none of the parent were available. Furthermore, when the diversity panel was used to estimate the marker effects, the prediction accuracies were negatively affected (Windhausen *et al.* 2012). Few possible reasons for decrease in accuracy of genomic prediction for across-within scenario include marker x population interaction, epistasis and different linkage phases between maker and QTL among populations (Schulz-Streeck *et al.* 2012). In an effort to enhance prediction accuracy, models including population specific marker effects (Schulz-Streeck *et al.* 2012) or only the preselected markers having low marker x genetic background interaction (Zhao *et al.* 2012) were investigated. However, no improvement in the prediction accuracy was observed. In a different scenario, when the estimation and prediction were performed across the bi-parental populations, the prediction accuracies were higher compared to within biparental population prediction (Albrecht *et al.* 2011; Schulz-Streeck *et al.* 2012; Zhao *et al.* 2012; Windhausen *et al.* 2012). The increase in prediction accuracy resulted from differences in mean performances of populations rather than kinship between estimation and prediction set and LD between markers and QTLs (Windhausen *et al.* 2012). As genetic variation among populations can be efficiently exploited through parental selection, GS application is not needed in this scenario. Albrecht *et al.* (2014) explored the possibility of topcross prediction across testers wherein disappointingly low accuracies were observed.

The prospects of GS for single-cross performance have been investigated with simulation and experimental studies in maize (Technow *et al.* 2012; Massman, *et al.*

2013a; Technow *et al.* 2014). All the studies reported high accuracies of genomic prediction of single crosses. Increasing the number of tested parents (0, 1, 2) (Technow *et al.* 2012; Massman, *et al.* 2013a; Technow *et al.* 2014) as well as increasing the number of single crosses per tested parent (Technow *et al.* 2014) significantly improved the prediction accuracies. Also, Technow *et al.* (2012) found small benefit of increasing the marker density and modelling population specific marker effects and dominance in the prediction model. The accuracies of GBLUP and Bayes B were very similar (Technow *et al.* 2014). In a comparison of genome and transcriptome-based single cross prediction, Zenke-Philippi *et al.* (2016) observed similar prediction accuracies of ridge regression model employing these two different type of markers.

The potential of genomic prediction of single-cross performance was also studied in other crops including wheat and rice in which moderate to high accuracies were observed. In a wheat dataset consisting of 90 single crosses derived from 22 females and 13 males elite lines, Zhao *et al.* (2013) investigated the predictive performances of RR-BLUP, Bayes A, Bayes B, Bayes C and Bayes Cπ models incorporating additive and dominance marker effects. The prediction accuracies were high (0.58-0.63) for all the models with slight superiority of RR-BLUP and Bayes B. In their study, ignoring the dominance effect resulted in equal or slightly higher prediction accuracies. In another largest experimental study in wheat, Zhao *et al.* (2015) performed genome and metabolite-based prediction of single-cross performance using 1604 single crosses generated by crossing 120 diverse female and 15 male lines. The mean genome-based prediction accuracies were 0.89 for T2 single crosses, 0.65 for T1 single crosses and 0.32

for T0 single crosses. They found no improvement in the prediction accuracy with either modelling epistasis or metabolite profiling. Xu *et al.* (2014) compared GBLUP, Bayes B and LASSO for predicting single-cross performance in rice. They used 278 single crosses generated between 210 RILs derived from single biparental population. All the three methods provided similar results. The predictabilities (squared correlation between observed and predicted values) for grain yield, number of tillers per plant, number of grain per panicle and 1000 grain weight were ranged from 0.09-0.16, 0.20-0.23, 0.35-0.37 and 0.67-0.69 respectively. Further, they found no noticeable improvement in prediction accuracy by including dominance and epistatic effects in GBLUP model.

OBJECTIVES

As described in the introduction, the early stages of hybrid breeding consist of generation of RILs or DHLs from several biparental population for evaluation in hybrid performance. The initial selection of lines is based on per se performance and topcross test using one or multiple testers and evaluation in single-cross combinations is delayed until advanced stages. This process has the advantage that lines having poor potential for hybrid performance are discarded in the early stages which allows concentration of resources on more promising lines. In accordance with this process of commercial hybrid development, published genomic hybrid prediction studies have investigated the potential of genomic prediction for topcross performance using the experimental material resembling to the early stages (Lorenzana and Bernardo 2009; Albrecht *et al.* 2011; Zhao *et al.* 2012; Jacobson *et al.* 2014) and single-cross performance using the experimental

material resembling to advanced stages of hybrid breeding (Massman, *et al.* 2013a; Technow *et al.* 2014). The current procedure of hybrid development, however, has some limitations which include more time for commercial hybrid development due to additional generations of topcross testing and inability to evaluate all possible single-cross combinations among the available lines which leaves open the possibility for losing some unique potential single crosses. Therefore, it would be desirable to investigate the potential of genomic prediction of single crosses in the early stages of hybrid development (Figure 2).

Also, published studies on genomic prediction of single-cross performance have mainly used parametric models such as GBLUP and Bayes A, Bayes B, Bayes C and Bayes Cπ. The parametric models make on a priori assumptions about the form of relationship between markers and genotypic value. These assumptions often do not hold in typical breeding populations limiting the ability of these models to precisely capture nonaddtive genetic effects (Gianola *et al.* 2006; Howard *et al.* 2014). In an alternate approach, nonparametric models for GS have been suggested. These models do not make prior assumption about the functional form between markers and phenotype. Rather they focus on prediction and seek a form that best fits to the TRS data. Simulation and experimental studies have showed better predictive performance of nonparametric models compared to parametric models for the traits conditioned by significant nonadditive genetic effects (Heslot *et al.* 2012; Crossa *et al.* 2013; Howard *et al.* 2014). Hybrid performance depends on both GCA of parents and SCA of cross. GCA is a function of average effects of genes while SCA is due to nonadditive i.e. both dominance

and epistatic genetic effects. It would therefore be desirable to investigate the potential of nonparametric GS models for predicting the single-cross performance.

Finally, very limited information is available about the optimizing TRS for genomic prediction of single-crosses (Technow *et al.* 2014). Previous studies on genomic prediction of single-cross performance highlighted the different criteria for TRS construction. These criteria included number of tested parents of a single crosses (i.e. 2, 1 and 0) and number of single crosses per tested parent. The results indicated that prediction accuracies increases with increase in the number of tested parents (Technow *et al.* 2012; Massman *et al.* 2013a; Technow *et al.* 2014) and number of single crosses per tested parent (Technow *et al.* 2014). Nevertheless, detailed and ready to use information on how to select the single crosses for phenotyping is lacking.

The overall goal of this study was to optimize the application of GS for prediction of single-cross performance. The three specific objectives decided for the present study were to:

1. Examine the potential of genomic prediction of single crosses in the early stages of hybrid development pipeline
2. Evaluate the nonparametric models for genomic prediction of early-stage single crosses
3. Optimize the training set composition for genomic prediction for early-stage single-crosses

Table 1. Commonly used parametric and non-parametric models of genomic selection

Model	Main features	Software packages	Reference
Parametric models	Assume certain form relationship between markers and genotypic value. Use shrinkage and/or variable selection procedure to estimate marker effects.		de los Campos *et al.* (2013)†
RRBLUP	1. Marker effects are assumed random having a normal distribution with common variance 2. Equally shrinks marker effects towards zero. No variable selection (i.e. None marker effect is zero). Penalty parameter is the ratio of residual variance (Ve) and common marker effect variance (V_β) 3. Robust compared to other models. Ideal for traits with many small effect QTLs.	rrBLUP, BGLR, ASReml-R	Meuwissen *et al.* (2001); Piepho (2009)
GBLUP	1. Numerator relationship matrix in BLUP is replaced by genomic relationship matrix estimated from genomic marker data 2. Computationally efficient and mathematically equivalent to RRBLUP 3. Genomic and pedigree relationship information can be combined in a single step method	rrBLUP, BGLR, ASReml-R	VanRaden (2008)
LASSO	1. Combines shrinkage and variable selection 2. Penalty is proportional to sum of marker effects i.e. L1 norm 3. Cannot select more variables (p) than sample size (n) when p >> n 4. Unstable with high dimensional data	glmnet	Li and Sillanpää (2012); Ogutu *et al.* (2012)
EN	1. Combines shrinkage and variable selection 2. Penalty is a weighted average of L1 and L2 norm 3. Robust to highly correlated predictors 4. Can selects more variables than sample size when p>> n	glmnet	Li and Sillanpää (2012); Zou and Hastie (2005)
BayesA	1. Applies only shrinkage and no variable selection 2. Marker specific variances are fitted. Prior for marker variances is scaled inverted chi-square distribution and prior for marker effects is scaled t distribution 3. Strong shrinkage of smaller marker effects towards zero and less shrinkage of sizable marker effects	BGLR GenSel	Meuwissen *et al.* (2001); Gianola (2013)
BayesB	1. Applies both shrinkage and variable selection 2. Marker specific variances are fitted. A proportion π of marker is assumed to have zero effect and remaining (1-π) markers variances assigned prior similar to BayesA 3. Suitable for traits with few large effects QTLs	BGLR GenSel	Meuwissen *et al.* (2001); Gianola (2013)

BayesC	1. Applies both shrinkage and variable selection 2. Marker specific variances are fitted. A proportion π of marker is assumed to have zero effect and remaining (1-π) markers variances assigned normal distribution prior	BGLR GenSel	De Los Campos *et al.* (2009)
BayesCπ	1. Applies both shrinkage and variable selection 2. Common marker variance is assumed and value of π is considered as unknown. Prior for marker variance is inverted chi-square. For π, prior is uniform (0, 1) distribution. 3. When π = 0, it is identical to RR-BLUP 4. Short computational time	GenSel	Habier *et al.* (2011)
Non-parametric models	Do not make strong assumption about the form of relationship between markers and genotypic value. They seek a form that best fits the training data while maintaining generality for new data. Thus, their main focus is on prediction. These methods are expected to capture nonadditive effects without explicitly modelling them.		González-Recio *et al.* (2014)[ξ]
RKHS	1. Genomic relationship matrix is replaced by kernel matrix that creates similarities among individuals 2. Gaussian kernel is typically used to define relationship between individuals 3. Equal or better predictive performance compared to parametric methods	BGLR	Gianola *et al.* (2006); Gianola and van Kaam (2008)
SVR	1. Like RKHS, use nonparametric kernel e.g. Gaussian radial basis kernel 2. Unlike RKHS which uses quadratic loss function, epsilon insensitivity loss function is used	Kernlab LIBSVM	Maenhout *et al.* (2007); Long *et al.* (2011)
NN	1. Consist of many processing units (i.e. neurons) which acts in parallel 2. Potential to capture complex relationship between the input and response variable 3. Susceptible to overfitting	brnn MATLAB	Gianola *et al.* (2011)

Abbreviations: RRBLUP - Ridge regression best linear unbiased prediction; GBLUP - Genomic best linear unbiased prediction; LASSO - Least absolute selection operator; EN - Elastic net; RKHS - Reproducing kernel Hilbert space; SVR - Support vector regression, NN - Neural network.

Software references: ASReml-R (Butler *et al.* 2009); BGLR (Pérez and de Los Campos 2014); brnn (Pérez-Rodriguez and Gianola 2013); glmnet (Friedman *et al.* 2010); GenSel (Fernando and Garrick 2008); kernlab (Zeileis *et al.* (2004); LIBSVM (Chang and Lin 2011); MATLAB (Demuth and Beale 2009); rrBLUP (Endelman 2011);

† - Review on parametric models of genomic selection

[ξ] - Review on non-parametric models of genomic selection

Table 2. Summary of published studies on genomic selection for per se performance, topcross performance and single-cross performance

Reference	Brief description	Experimental material	Model	Cross-validation	Prediction accuracy†
A. *Genomic selection for per se performance*					
Technow *et al.* (2013)	Accessed the prospects of genomic prediction of northern corn leaf blight resistance and combining inbred lines across heterotic groups into TRS	*Germplasm:*100 dent and 97 flint inbred lines *Markers:* 37908 SNPs	GBLUP	CV_WW* CV_AW$ CV_AAξ	0.33-0.64 (CV_WW) 0.08-0.3 (CV_AW) 0.37-0.71 (CV_AA)
Riedelsheimer *et al.* (2013)	Investigated the effect of different level of relatedness between TRS and TS on prediction accuracy within BP for two disease trait and three grain yield traits	*Germplasm*: 635 DH lines from the five interconnected BP *Markers*: 16741 SNPs	GBLUP	CV_WW CV_AW	0.59 (CV_WW), 0.05-0.34 (CV_AW)
B. *Genomic selection for topcross performance*					
Lorenzana and Bernardo (2009)	Compared the prediction accuracies of MLR, GBLUP and e-Bayes methods and evaluated the effect of TRS size and number of markers	*Germplasm*: Testcrosses of RIL/DHLs belonging to three BP *Markers*: 1339 SSR or RFLP; 125 SNPs	GBLUP e-Bayes	CV_WW	0.25-0.64
Albrecht *et al.* (2011)	Examined the accuracies of within versus across family prediction. Also assessed the effect of TRS size and different approaches of estimating genetic relationship.	*Germplasm*: Testcrosses of 1380 DH lines from 36 BP belonging to dent heterotic group *Markers*: 1152 SNPs	GBLUP	CV_WW CV_AW CV_AA	0.26-0.59 (CV_WW) 0.47-0.48 (CV_AW) 0.72-0.74 (CV_AA)
Riedelsheimer *et al.* (2012)	Investigated the usefulness of genome and metabolite-based prediction	*Germplasm:* testcrosses of 285 diverse inbred lines *Markers:* 56110 SNPs and 130 metabolites	RRBLUP	CV_AA	0.60-0.78

Schulz-Streeck *et al*. (2012)	Evaluated the advantage of modelling main and population specific marker effects. Also compared RRBLUP, ridge regression, LASSO and elastic net	*Germplasm*: Testcrosses of 312 DH lines from five BP *Markers*: 39339 SNPs	RRBLUP RR LASSO EN	CV_AW CV_AA	0.024-0.31 (CV_AW) 0.28-0.37 (CV_AA) Note: predictive ability (heritability not given)
Windhausen *et al.* (2012)	Evaluated the prospects of marker effects estimated in diversity panel for prediction within a biparental population	*Germplasm*: Testcrosses of 255 inbreds from diversity panel and 150 inbreds belonging to 5 BP *Markers*: 18695 SNPs	GBLUP	CV_AW CV_AA CV_AW_{group} CV_AA_{group}	-0.42-0.37 (CV_AW) 0.46-0.54 (CV_AA) 0.14-0.26 (CV_AW_{group}) 0.15-0.39 (CV_AA_{group})
Zhao *et al.* (2012)	Compared the prediction within and across biparental families. Also, evaluated the effect of modelling preselected markers with low genetic background interaction effect.	*Germplasm*: Testcrosses 788 $F_{3:4}$ lines from six BP *Markers*: 960 SNPs	GBLUP	CV_WW CV_AW CV_AA	0.40-0.64 (CV_WW) 0.39-0.70 (CV_AW) 0.45-0.69 (CV_AA)
Crossa *et al.* (2013)	Compared the different methods of incorporating genotyping by sequencing (GBS) marker data for genomic prediction with GBLUP and RKHS	*Germplasm:* Testcrosses 505 DH lines and diverse panel of 296 maize inbred lines *Markers:* GBS	GBLUP RKHS	CV_AA	0.60-0.90
Massman, *et al.* (2013b)	Assessed the usefulness of marker effects estimated from single cross data for test cross prediction	*Germplasm:* Testcrosses of 5 BP along with 479 single crosses between 59 BSSS inbreds and 49 NSSS inbreds Markers: 669 SNPs	GBLUP RRBLUP	CV_AW	-0.08 – 0.36
Albrecht *et al.* (2014)	Accessed the efficiency of prediction across genetic groups and tester. Also compared the potential of predicting across locations and across years	*Germplasm*: Testcrosses of 1,073 and 857 DH lines derived from multiple biparental families *Markers*: 56110 SNPS	GBLUP	CV_WW_{group} CV_AW_{group} $CV\_AW_{group/tester}$ CV_AA_{group}	0.36-0.77 (CV_WW_{group}) 0.31-0.35 (CV_AW_{group}) 0.14-0.53 ($CV\_AW_{group/tester}$) 0.45-0.74 (CV_AA_{group})

Jacobson *et al.* (2014)	Evaluated the usefulness of GCA model for genomewide selection within a BP	*Germplasm:* Tetscrosses of 970 BP *Markers:* 49 to 100 SNPs	RRBLUP	CV_AW CV_WW	-0.16-0.63 (CV_WW) 0.02-0.65 (CV_AW)
C. *Genomic selection for single-cross performance*					
Maenhout *et al.* (2007)	Compared SVR and GBLUP for prediction of single-cross performance	*Germplasm:* 2371 single crosses between 105 BSSS and 93 Iodent lines *Markers:* 75 SSR and AFLP	SVR GBLUP	LOOCV	0.66
Massman, *et al.* (2013a)	Compared BLUP with RRBLUP for single-cross prediction	*Germplasm*: 479 single crosses between 59 BSSS inbreds and 49 NSSS inbreds *Markers*: 669 SNPs	BLUP RRBLUP	*k*-fold CV for T2, T1 single crosses	0.87 (T2) 0.73-0.75 (T1)
Technow *et al.* (2014)	Evaluated the prospects of single cross prediction using GBLUP and BayesB	*Germplasm*: 1254 single crosses between 123 dent and 86 flint inbred lines *Markers*: 35478 SNPs	GBLUP BayesB	*k*-fold CV for T2, T1 and T0 single crosses	0.86-0.92 (T2) 0.82-0.86(T1) 0.75-0.78(T0)

Different cross validation scenarios: *Training set (TRS) and test set (TS) sampled within a biparental population; $ TRS sampled across biparental populations and TS sampled within a biparental population; ξ TRS and TS sampled across biparental populations. Subscripts group and group/tester are used to denote above three cross-validation scenarios with reference to group and group/tester instead of biparental population.

† Prediction accuracies for grain yield (unless specified) when TRS and TS were evaluated in the same environment/s or across environments.

Acronyms: BP - Biparental populations; GCA - General combining ability; BSSS - Iowa Stiff Stalk Synthetic; LOOCV - Leave-one-out cross-validation; NSSS - Non-Stiff Stalk Synthetic; e-Bayes - empirical Bayes; GBLUP - Genomic best linear unbiased prediction; MLR - Multiple linear regression; RRBLUP - Ridge regression best linear unbiased prediction; RKHS - Reproducing kernel Hilbert spaces.

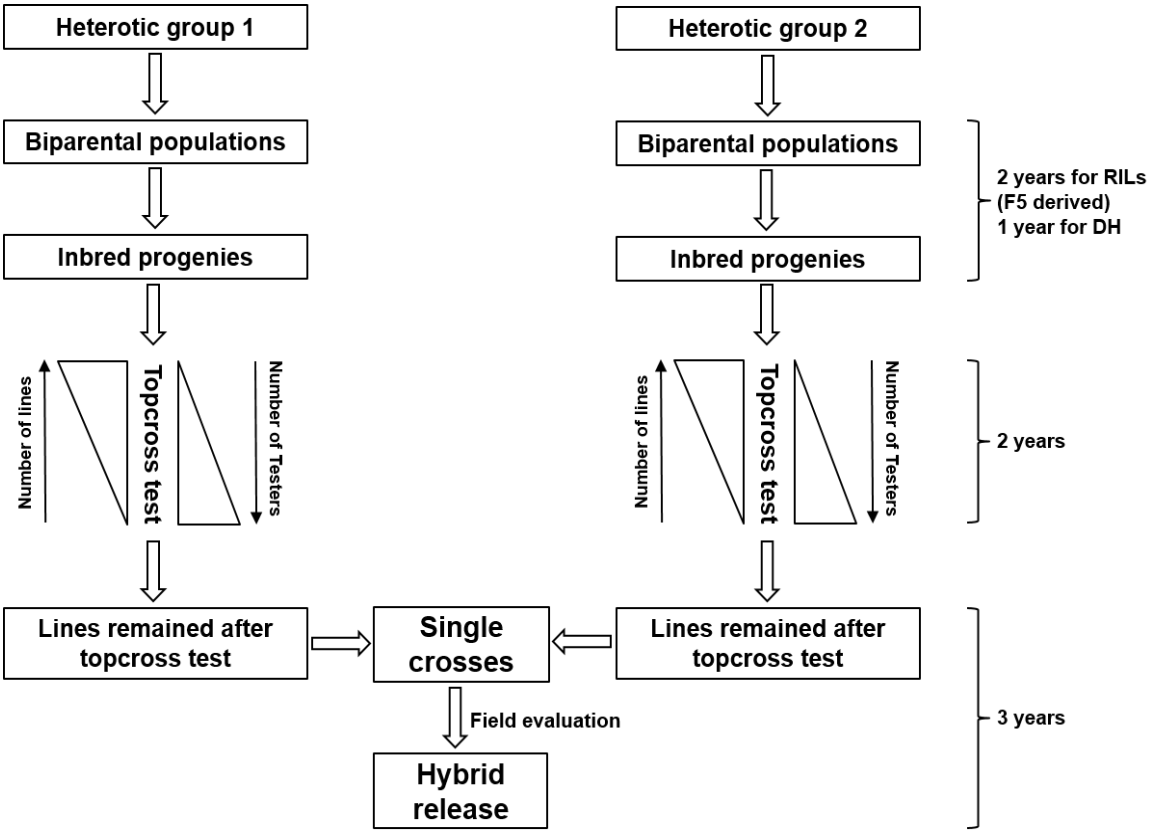


Figure 1. Schematic outline of typical hybrid maize breeding pipeline. Estimated timeline for various stages adapted from Heffner *et al.*, (2010).

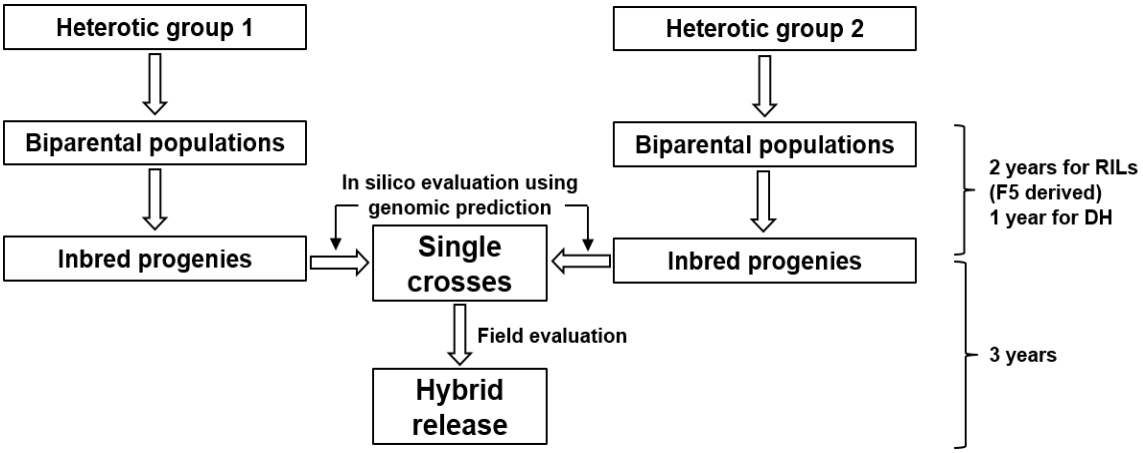


Figure 2. Schematic outline of hybrid maize breeding pipeline with genomic selection. Estimated timeline for various stages adapted from Heffner *et al.*, (2010). This scheme assumes the use of historical data for model training.

Chapter 2: Genomic Prediction of Single Crosses in the Early Stages of a Maize Hybrid Breeding Pipeline

Prediction of single-cross performance has been a major goal of plant breeders since the beginning of hybrid breeding. Recently, genomic prediction has shown to be a promising approach, but only limited studies have examined the accuracy of predicting single-cross performance. Moreover, no studies have examined the potential of predicting single crosses among random inbreds derived from a series of biparental families, which resembles the structure of germplasm comprising the initial stages of a hybrid maize breeding pipeline. The main objectives of this study were to evaluate the potential of genomic prediction for identifying superior single crosses early in the hybrid breeding pipeline and optimize its application. To accomplish these objectives, we designed and analyzed a novel population of single crosses representing the Iowa Stiff Stalk Synthetic/Non-Stiff Stalk heterotic pattern commonly used in the development of North American commercial maize hybrids. The performance of single crosses was predicted based on parent's combining ability and covariance among single crosses. The prediction accuracies estimated using cross-validation ranged from 0.39 to 0.77 for grain yield, 0.72 to 0.92 for plant height and 0.57 to 0.94 for staygreen depending on the number of tested parents of the single cross and genomic prediction method used. The prediction accuracies based on genomic estimated general and specific combining abilities were similar to those based on genomic covariances among single crosses. Overall, our results suggest that genomic prediction of single crosses in the early stages of a hybrid breeding pipeline holds great potential to re-design hybrid breeding and increase its efficiency.

INTRODUCTION

Contemporary hybrid breeding programs are based on the 'pure-line method of corn breeding' proposed by Shull (1909). This method includes the development of inbreds by self-pollination followed by evaluation of selected inbreds for single-cross performance when crossed to other inbreds. A major challenge with this method is achieving adequate testing of the inbreds to evaluate performance in single-cross combinations (Hallauer *et al.* 1988). In maize, heterotic groups are well defined, and single crosses are almost exclusively made between heterotic groups. The fullest assessment of single-cross performance in maize, therefore, would be a complete factorial mating design achieved by making all between-heterotic group single crosses. This would provide complete information on both general combining ability (GCA) and specific combining ability (SCA) (Comstock and Robinson 1948). However, a full factorial among inbreds can be cost prohibitive as advanced hybrid breeding programs typically have many inbreds to evaluate, making the number of all possible single crosses extremely large. For this reason, predicting single-cross performance has always been a major issue in all hybrid breeding programs (Schrag *et al.* 2009).

Several approaches have been used to evaluate the genetic merit of inbreds for single-cross performance with variable success. These approaches include inbred per se performance, performance when crossed to testers ("topcross" test), best linear unbiased prediction (BLUP) using pedigrees, and molecular marker-assisted prediction. Many of these approaches have been reviewed in detail elsewhere (Smith 2004; Schrag *et al.* 2009). Per se performance of inbred is typically found to be a very poor predictor of

single-cross performance, especially for traits such as grain yield, where strong dominance effects underlie the genetic variance (Love and Wentz 1914; Hallauer 1977; Smith 1986). A topcross test is an established and simple approach to assess the genetic worth of inbreds in single-cross combinations (Jenkins and Brunson 1932). However, topcross evaluation of a large number of inbreds is difficult (Albrecht *et al.* 2011) and selections based on single-cross performances are carried out in later stages which increases the time required for commercial hybrid development. (Bernardo 1996a) showed that pedigree-based BLUP is useful for prediction of untested single crosses. He used pedigree-based covariance matrices among tested and untested single crosses to obtain BLUPs for untested single crosses. The correlations between observed and predicted performance were moderate (0.43-0.76) for single crosses whose both parents were tested in single-cross combinations. However, when one or both of the parents of the single cross were untested, the correlations were severely decreased (Bernardo 1996c).

The relationship between genetic distance (GD) of parental inbreds, measured by molecular markers, and heterosis has been extensively studied in maize. While it is possible to predict single-cross performance using marker based GD for hybrid sets composed of both intra- and inter-heterotic group single crosses, correlations for predicting inter-heterotic group single crosses only were reported to be very low (Melchinger 1999; Lee *et al.* 2007). Two possible causes of these low prediction accuracies include (1) loose association between heterotic QTL and the molecular markers used to estimate GD and (2) opposite linkage phases between the QTL and

marker alleles as generally expected with inter-heterotic single crosses (Charcosset *et al.* 1991; Bernardo 1992). Commercial hybrids consist of only inter-heterotic group single crosses, making them the only ones relevant for prediction in breeding programs. In a modified approach, prediction of single-cross performance and SCA based on only significant markers was suggested (Vuylsteke *et al.* 2000), but this approach was found to be inferior to an established GCA method. Also, extending the GCA predictions with SCA estimates from associated markers did not improve the prediction accuracy (Schrag *et al.* 2006, 2007).

Genomic prediction is an approach that uses markers to predict the genetic value of complex traits in progeny for selection and breeding (Meuwissen *et al.* 2001). When genomic predictions are used to make selections, it is referred to as genomic selection (GS). The primary difference between GS and traditional forms of marker-assisted selection (MAS) is the simultaneous use of a large number of markers distributed genome-wide as opposed to a small set of markers linked to QTL (Heffner *et al.* 2009). Implementation of genomic prediction and selection requires the development of training or calibration sets consisting of individuals that have been both phenotyped and genotyped, followed by model calibration. A whole suite of genomic prediction models have been developed, each deploying different strategies to estimate genome-wide marker effects (de los Campos *et al.* 2013).

Recently, published results from simulation and experimental studies have given first indications of usefulness of genomic prediction models for hybrid performance in

maize (Albrecht *et al.* 2011; Technow *et al.* 2012; Windhausen *et al.* 2012; Massman, *et al.* 2013a; Jacobson *et al.* 2014; Albrecht *et al.* 2014; Technow *et al.* 2014). However, most of the experimental studies were focused on prediction of topcross performance using single tester (Albrecht *et al.* 2011; Windhausen *et al.* 2012; Jacobson *et al.* 2014; Albrecht *et al.* 2014). Experimental studies on genomic prediction of single-cross performance have been based on historical data consisting of established inbred parents with mixed and complex ancestry (Massman, *et al.* 2013a; Technow *et al.* 2014). These studies used covariances among tested and untested single crosses estimated from realized genomic relationship matrices to predict the performance of untested single crosses. The prediction accuracies were high, often exceeding 0.75, even when both parents of the single cross were untested.

Identification of superior single crosses early in the hybrid breeding pipeline would be beneficial to develop commercial hybrids more quickly. The current practice of initial selection among available inbreds based on their topcross performance followed by evaluation of single crosses made among selected inbreds increases time required for commercial hybrid development. Moreover, not all possible single-cross combinations among available inbreds gets evaluated with this approach. It is important, therefore, to study the potential of genomic prediction of early-stage single crosses i.e. single crosses between random set of inbreds from each heterotic group skipping topcross test based inbred selection. With this in mind, the main objective of this study was to evaluate the potential of genomic prediction for identifying superior single crosses early in the breeding pipeline. Also, we evaluated how the prediction model and the composition of

the training set affected the single-cross prediction accuracy. To accomplish these objectives, we designed and analyzed a novel population of single crosses. The parental recombinant inbred lines (RILs) and doubled haploid lines (DHLs) were randomly selected from three Iowa Stiff Stalk Synthetic (BSSS) and three Non-Stiff Stalk Synthetic (NSSS) biparental populations. All single crosses, therefore, represented the BSSS/NSSS heterotic pattern commonly used in the development of North American commercial maize hybrids. All RILs and DHLs were genotyped using genotyping by sequencing (GBS) (Elshire *et al.* 2011), which represents an affordable genotyping option that is critical to the routine use of these methods in a breeding program.

MATERIALS AND METHODS

Germplasm

Three BSSS inbred parents (PHG39, PHJ40, and B73) and three NSSS inbred parents (LH82, PHG47, and PHG84) were used for creating six biparental families by making each of the three possible crosses between the three BSSS inbreds and also between the three NSSS inbreds. The chosen parents were identified as being both genetically diverse and superior in GCA for grain yield under high planting density (Mansfield and Mumm 2014). A total of 217 lines were developed from crosses between these parents. Approximately 10% of these lines were RILs and 90% DHLs. RILs and DHLs will be hereafter referred collectively as "inbred progenies". The number of inbred progenies in each of the six biparental families ranged from 2 to 69 (Table 3). Random crosses among the inbred progenies were made between heterotic groups to produce 312

single crosses (Figure 3). Single crosses representing each biparental family were balanced to the extent possible while maximizing the number of inbred progenies used in the single crosses. Completely balanced representation was not achieved due to seed limitations and comparatively fewer inbred progenies available for certain biparental families. Single crosses were grouped into nine "single-cross families", which we defined as a group of single crosses created using inbred progenies from the same biparental family on each side of the heterotic pattern (Table 1). For example, a single cross with pedigree (PHJ40×PHG39)DH-1/(PHG47×PHG84)DH-1 belongs to the same single-cross family as a single cross with pedigree (PHJ40×PHG39)DH-2/(PHG47×PHG84)DH-2. The mean number of times an individual BSSS inbred progeny was used in a cross was 6.9. The mean number of times an individual NSSS inbred progeny was used in a cross was 1.8. Number of single crosses per single-cross family ranged from 19 to 51 (Table 3).

Field Experiments

The 312 single crosses were evaluated at two locations in 2012 and three locations in 2013. Two locations were common between years. The locations were as follows: South Farms (Urbana, IL; 2012 & 2013), Maxwell Farms (Urbana, IL; 2012 & 2013) and Monmouth (IL; 2013 only). The five location–year combinations were defined as separate environments. The experimental design was an $\alpha(0, 1)$-incomplete block design (Patterson and Williams 1976) with three replications at each environment. All trials were planted with an Almaco Seed Pro 360 planter set at 0.64 m row spacing and 4.46 m long row. Entries were grown in small plots consisting of two rows. Plots were

overplanted by 15% to compensate for germination failure and later thinned to the target plant density of 116,000 plants ha^{-1}. All fields were controlled for weeds. Nitrogen (N) was applied before planting as 28% urea-ammonium nitrate at a rate of 336.4 kg ha^{-1} to all fields. Phosphorous and potassium were each applied at 112 kg ha^{-1} according to recommended levels determined by soil tests performed by the University of Illinois Crop Science Research and Education Center. Stand counts were recorded and plots with planting densities lower than 106,000 plants ha^{-1} discarded. Additionally, issues with seed production result in fewer single crosses being planted at all locations in 2013 (South Farms: 260; Maxwell Farms: 259 & Monmouth: 258). Plots were machine harvested and data were recorded for several agronomic traits. For this study, data on grain yield (GY), plant height (PH) and staygreen (SG) were used for downstream analyses. GY was converted to Mt ha^{-1} on a 155 g kg^{-1} moisture basis. PH was measured post anthesis on a single representative plant determined by visually surveying the entire plot before measurement. SG was evaluated visually as a percentage of total dry down, where a rating of 1 represented complete senescence and a rating of 10 represented fully green leaves.

Genotyping by Sequencing

Five plants of each inbred progenies were germinated. A total of 0.1 g of tissue was sampled from leaf tips and pooled across the five plants. DNA was extracted using the Qiagen DNeasy Plant 96 kit following the DNeasy Plant Handbook. DNA samples were sent to the Institute for Genomic Diversity (IGD) at Cornell University for

genotyping by sequencing (GBS) where library construction and sequencing was performed as described by (Elshire *et al.* 2011). Single nucleotide polymorphisms (SNPs) were scored from the raw sequence data using the TASSEL GBS Pipeline version 3.0 (Glaubitz *et al.* 2014). SNPs with greater than 20% missing values and less than 5% minor-allele frequency were removed from the dataset. Heterozygotes were treated as missing data. Missing data was replaced by the mean value for the markers (i.e. naïve imputation). Of the markers remaining after filtration, markers that were polymorphic among both BSSS and NSSS inbred progenies were retained for analysis. The final marker data set consisted of 2296 high-quality SNPs.

Phenotypic Data Analysis

The phenotypic data were unbalanced due to missing observations. We used the following statistical model for the analysis of the data across the five environments

$$y_{iklq} = \mu + g_i + e_k + (ge)_{ik} + r_{l(k)} + b_{q(kl)} + \varepsilon_{iklq} \quad \ldots\ldots (1)$$

where y_{iklq} is the phenotypic observation for i^{th} single cross evaluated in the k^{th} environment in the l^{th} complete block (i.e. replicate) and q^{th} incomplete block. The effects in the model are as follows: μ is the grand mean; g_i represents effect of the i^{th} single cross; e_k represents the effect of the k^{th} environment; $(ge)_{ik}$ represents the interaction effect between single cross and environment; $r_{l(k)}$ represents the effect of the l^{th} complete block nested within the k^{th} environment; $b_{q(kl)}$ represents the effect of the q^{th} incomplete block nested within the l^{th} complete block in the k^{th} environment; and ε_{iklq}

represents the residual. Environment and replication nested within environment effects were modeled as fixed effects while all other effects were treated as random. The distribution of g_i was as follow: $g_i \sim N(0, I\sigma^2)$. Error and incomplete block variances were allowed to be heterogeneous among environments.

The above model was implemented using ASReml-R software (Butler *et al.* 2009) to obtain restricted maximum likelihood estimates (REML) of all variance components and solve the mixed linear model equations. Significance of the variance components was determined using likelihood ratio tests at 0.001 level of significance. The entry-mean heritability of each trait was computed according to (Holland *et al.* 2003) as: $H^2 = \sigma_g^2 / \left(\sigma_g^2 + \frac{\sigma_{g\times e}^2}{h_k} + \frac{\sigma_e^2}{h_t}\right)$, where, $\boldsymbol{\sigma_g^2}$ represents the variance among single crosses, $\boldsymbol{\sigma_{g\times e}^2}$ represents the variance of interaction effects of single crosses with environments, $\boldsymbol{\sigma_e^2}$ is the residual variance, h_k is the harmonic mean of number of observations per single cross within an environment, and h_t is the harmonic mean of total number of observations per single cross. Similarly, model (1) used to estimate the genetic variance and broad sense heritability for individual single-cross family. Finally, we calculated best linear unbiased predictions (BLUP) of single crosses and used these to evaluate single-cross prediction accuracy in further analyses.

Single-Cross Prediction Methods

The linear model used for single-cross performance was

$$y_{ijklq} = \mu + f_i + m_j + s_{ij} + e_k + r_{l(k)} + b_{q(kl)} + (fe)_{ik} + (me)_{jk} + (se)_{ijk} + \varepsilon_{ijklq} \quad ...(2)$$

where y_{ijklq} is the phenotypic observation on a single cross between the i^{th} and j^{th} inbred progeny evaluated in the k^{th} environment in the l^{th} complete block and q^{th} incomplete block. The effects in the model are as follows: μ is the grand mean; f_i and m_j represents the GCA effects of the female (BSSS inbred progenies) and males (NSSS inbred progenies), respectively; s_{ij} represents the SCA effect of the single cross; $(fe)_{ik}$, $(me)_{jk}$, and $(se)_{ijk}$ represent the interaction effects of respective terms with the k^{th} environment. The remaining terms were as described in the model (1).

The random effect vectors f, m, and s were assumed to have the following multivariate normal distributions: $f \sim MVN\left(0, G_f \sigma^2_{GCA\_F}\right)$, $m \sim MVN\left(0, G_m \sigma^2_{GCA\_M}\right)$, $s \sim MVN(0, S\sigma^2_{SCA})$, where G_f and G_m were additive genomic relationship matrices of females and males, respectively, calculated according to Method 1 of VanRaden (2008). The dominance relationship matrix, S, was computed according to (Bernardo 2002) using the corresponding elements from matrices G_f and G_m. The above model (2) was implemented using ASReml-R software (Butler *et al.* 2009).

We evaluated four methods to predict single-cross performance using the model (2). Broadly, these methods can be grouped into two categories: 1. Parent GCA and SCA effects; 2. Additive and dominance covariances among single crosses.

1a. Parent GCA

Performance of untested single crosses $(\hat{y}_u)$ was predicted from the GCA of the corresponding parents, *i* and *j* estimated from model (2) as

$$\hat{y}_u = \hat{\mu} + \hat{f}_i + \hat{m}_j \quad \text{...(3)}$$

GCA of females and/or males with no performance data of their single crosses were estimated from related inbred progenies using the additive genomic relationship matrix in model (2).

1b. Parent GCA plus single-cross SCA

Performance of untested single crosses $(\hat{y}_u)$ was predicted using sum of the parent GCA and SCA of the single crosses as

$$\hat{y}_u = \hat{\mu} + \hat{f}_i + \hat{m}_j + \hat{s}_{ij} \quad \text{…(4)}$$

Like the GCA effects, the SCA effects for untested single crosses were estimated using the dominance genomic relationship matrix in model (2).

2a. Additive covariance among single crosses

The performance of untested single crosses $(\hat{y}_u)$ was predicted based on the covariance among tested and untested single crosses as

$$\hat{y}_u = C_{ut} C_{tt}^{-1} y_t \quad \text{…(5)}$$

Where, C_{ut} is the genetic covariance matrix of untested and tested single crosses, C_{tt} is the phenotypic covariance matrix of the tested single crosses and y_t is a vector of tested single-cross BLUPs obtained from model (1). The elements of C_{ut} and C_{tt} were computed according to (Bernardo 2002) using the genomic relationship matrices G_f and G_m. Briefly, let i and i' denote any two female inbred progenies and j and j' any two

male inbred progenies. For a given pair of single crosses, ($i \times j$) and ($i' \times j'$), the elements of C_{ut} and the off diagonal elements of C_{tt} were calculated as $(G_f)_{ii'}\ \sigma^2_{GCA\_F}$ + $(G_m)_{jj'}\sigma^2_{GCA\_M}$. The diagonal elements of C_{tt} were estimated as $(G_f)_{ii}\ \sigma^2_{GCA\_F}$ + $(G_m)_{jj}\sigma^2_{GCA\_M} + \sigma^2_{\bar{X}}$ where $\sigma^2_{\bar{X}}$ was equal to σ^2_e divided by the total number of observations for single cross $(i \times j)$. The estimates of $\sigma^2_{GCA\_F}$ and $\sigma^2_{GCA\_M}$ were obtained from model (2).

2b. Additive plus dominance covariance among single crosses

The method described in 2a was extended by including dominance covariance among the tested and untested single crosses. Specifically, the elements of C_{ut} and off diagonal elements of C_{tt} were computed as $(G_f)_{ii'}\sigma^2_{GCA\_F}$ + $(G_m)_{jj'}\sigma^2_{GCA\_M} + (G_f)_{ii'}(G_m)_{jj'}\sigma^2_{SCA}$. The diagonal elements of C_{tt} were estimated as $(G_f)_{ii}\ \sigma^2_{GCA\_F}$ + $(G_m)_{jj}\sigma^2_{GCA\_M} + (G_f)_{ii}(G_m)_{jj}\sigma^2_{SCA} + \sigma^2_{\bar{X}}$. The estimates of σ^2_{SCA} was obtained from model (2).

Cross-Validation and Prediction Accuracy Estimation

Accuracy of single-cross prediction was evaluated using leave-one-out cross-validation (LOOCV). LOOCV is a particular case of *k*-fold cross validation with *k* = *n*. We chose LOOCV because the greater number of folds minimizes bias in the estimator (Kohavi 1995). Five different LOOCV scenarios involving varying degrees of relationship between training and validation set single crosses were considered (Figure 4). The cross-validation scenarios were as follows: 1) T2 -- Both parents of a single cross

contained in the validation set were tested. 2) T1F -- Only the female parent of a single cross contained in the validation set was tested. 3) T1M -- Only the male parent of a single cross contained in the validation set was tested. 4) T0 -- Neither of the parents of a single cross contained in the validation set was tested. 5) Novel single-cross family -- All single crosses belonging to one single-cross family were removed from the training set and thus formed the validation set. The conventional LOOCV was slightly modified to maintain constant training set size for each of the five cross-validation scenarios considered. The common maximum possible training set size across the five scenarios was 261. With this in mind we decided to set the training set size to 250 for all the five cross-validation scenarios in order to remove the confounding effect of population size. For the first four scenarios, the cross validation is repeated such that each of the 312 single crosses was placed into the validation set exactly one time (i.e. leave-one-individual-out cross-validation). For each of the 312 rounds, a random sample of 250 single crosses from the remaining single crosses was drawn without replacement and formed the training set. This was repeated 30 times to allow for sufficient re-sampling of the training for a total of 9360 (30 x 312). For each of the 30 repetitions, the predictions were integrated into a single vector and correlated with the phenotypic observations as described below. For scenario 5, the cross validation was repeated so that each of the nine single-cross families was entered into the validation set one time (i.e. leave-one-family-out cross-validations). This was repeated 30 times by re-sampling 250 single crosses without replacement from the training set. The prediction accuracy, however, was

evaluated only for the six largest families because size of the three families (f7, f8, f9) was too small to accurately estimate correlation coefficients (Table 3).

The single-cross BLUPs estimated from the phenotypic data were treated as the observed single-cross performance and used as the basis to evaluate single-cross prediction methods. Prediction accuracy was expressed as the Pearson's correlation coefficient between the observed and predicted single-cross performance divided by the square root of the broad-sense heritability on an entry-mean basis (Dekkers 2007). The mean prediction accuracy across the 30 repetitions was reported. Standard errors of the prediction accuracy were calculated using the bootstrap method implemented in the R package *boot* (Canty 2014). Briefly, for each of the 30 repetitions, the predicted and observed values were resampled with replacement for 200 times. The distribution of 200 correlation coefficient estimates was used to estimate the bootstrap SE. This procedure was repeated for each of the 30 repetitions the mean standard error of 30 repetitions was reported.

RESULTS

Variance Components and Broad-Sense Heritability

Variance among single crosses (σ_g^2) was significantly different from zero $(\alpha = 0.001)$ in the whole population as well as within individual single-cross families for all three traits (Table 4). For GY, the entry-mean heritability was 0.58 across the whole population of single crosses and it ranged from 0.53 to 0.83 within the individual single-cross families. Similarly, for PH and SG, the entry-mean heritability was 0.89 and 0.81 in

the whole population, respectively, and ranged from 0.88 to 0.91 and 0.67 to 0.80 within individual single-cross families, respectively. The sum of parent σ^2_{GCA} was greater than σ^2_{SCA} for all traits. The proportion of σ^2_{SCA} was highest for GY, followed by PH and SG (Table 5).

Prediction Accuracy for T2, T1F, T1M and T0 Scenarios

We first evaluated the prediction accuracy for T2, T1F, T1M and T0 scenarios in the whole population using leave-one-individual-out cross-validation. Higher prediction accuracies were observed for SG and PH compared to GY for all scenarios (Figure 5). Prediction accuracies were highest for T2, followed by T1F, T1M and T0. The four methods provided similar prediction accuracies across different traits and scenarios.

Prediction Accuracy for Novel Single-Cross Family

We next investigated the potential to predict the performance of single crosses in a new single-cross family using the phenotypic and genotypic information on the single crosses from related single-cross families (leave-one-family-out cross-validation). When eight of the families were used as a training set to predict performances of single crosses within the remaining family, prediction accuracies were generally moderate for GY and high for PH and SG (Figure 6). The mean accuracies of methods 1a and 1b for prediction of novel single-cross families were, respectively, 0.61 and 0.61 for GY; 0.76 and 0.76 for PH; and 0.78 and 0.78 for SG. Variation in prediction accuracy across families was observed, especially for GY. We also evaluated the effect of adding single crosses from the family being predicted to the training set by comparing prediction accuracy of

individual family with leave-one-individual-out and leave-one-family-out cross-validations. The goal of this analysis was to measure the benefit of including information from the same single-cross family to accurately separate single crosses within the same family. Although the prediction accuracies were increased slightly for some families, they were decreased for other families (Figure 6). The mean prediction accuracies of methods 1a and 1b, respectively, were 0.65 and 0.61 for GY; 0.84 and 0.85 for PH; and 0.80 and 0.78 for SG.

Genomic Predictions of Grain Yield of All Possible Single Crosses

Genomic predictions were calculated for all possible 7866 single crosses between 46 BSSS and 171 NSSS inbred progenies based on the prediction model including parent GCA and cross SCA effects (i.e., Method 1b). The genomic predictions for GY ranged from 7.5 - 9.5 Mt ha^{-1}. The top 100 single crosses based on genomic predictions included only one single cross that was actually made and tested; the remaining 99 single crosses were never made. Moreover, more than 50 untested single-cross combinations surpassed the highest genomic prediction of any tested single cross (Figure 7).

DISCUSSION

Typical hybrid maize breeding programs involve the creation of large biparental families for topcrossing to elite testers early in the breeding pipeline. Early-stage selections are performed on the basis of topcross performance with a single elite tester, which is the sum of the candidate line GCA effect and any SCA effect between the candidate line and tester. While this is a very convenient and routine method, it has long

been recognized that it would be ideal to test all combinations of possible parents immediately in the hybrid breeding pipeline (Fehr 1987). There are two main advantages of early evaluation of all potential single crosses. First, it could identify the best parental combination immediately after progeny development. Selection of inbred progenies on the basis of topcross evaluation only leaves open the possibility that some unique parental combinations never made and evaluated could actually be commercially hybrids (Bernardo 2002). Secondly, early evaluation based on single-cross performance would enable the development of hybrids in shorter duration of time by essentially skipping the topcross test and immediately going to single-cross evaluation. Despite these advantages, field testing of all potential single crosses of inbred progenies is completely impractical for a mature hybrid maize breeding program.

Advances in genotyping technology, such as GBS, has made it very practical to genotype all parental candidate lines with dense, genome-wide markers (He *et al.* 2014). Genomic prediction models can predict the performance of all possible single-cross combinations, allowing the in-silico evaluation of all parental combinations just as in the ideal scenario. In the present study, GBS and yield trial data was used to build genomic prediction models for predicting single-cross performance. The single-cross prediction accuracies estimated using cross-validation ranged from 0.39 to 0.77 for grain yield, 0.72 to 0.92 for plant height and 0.57 to 0.94 for staygreen depending on the number of tested parents of the single crosses and genomic hybrid prediction method used. These prediction accuracies were 52-100, 76-98 and 64-100 percent of the estimated phenotypic accuracies $\left(\sqrt{h^2}\right)$ for GY, PH and SG respectively. The prediction accuracies of single-

cross performance achieved in this study, therefore, indicate that this approach holds great potential for increasing the efficiency of a hybrid breeding program by enabling the effective evaluation of all single-cross combinations.

Prediction Accuracy for T2, T1F, T1M and T0 Single Crosses

In order to understand the effect of tested versus untested parents, we evaluated the accuracies of prediction of single crosses having both (T2), either female (T1F) or male (T1M), or no (T0) parent tested for single-cross performance. Observed differences in prediction accuracies between these scenarios were considerable, with the highest prediction accuracy for T2 single crosses followed by T1 (T1F or T1M) and T0 single crosses. The T0 scenario was the most difficult to predict. Similar trends have been observed using simulations (Technow *et al.* 2012) as well as experimental studies based on historical data in maize (Massman, *et al.* 2013a; Technow *et al.* 2014), and wheat (Zhao *et al.* 2015). This finding can be explained by the representation of parents among a differing number of single-cross combinations in the training set. As the number of parents and single-cross combinations for each parent increases, the information shared between the single crosses being predicted and the training set increases. As a result, the GCA effects are estimated with high accuracy. In the T2 scenario, both parents are tested in multiple single-cross combinations within the training set, enabling accurate estimation of parent GCA effects. With a preponderance of GCA variance over SCA variance, genotypic values of T2 single crosses can, therefore, be predicted with higher accuracy. In the case of T1 scenario, however, only one of the parents is tested in single-cross

combination. Consequently, the prediction accuracy of T1 single crosses is lower than for T2 single crosses. In the present study, the prediction accuracy of the T1F single crosses is greater than that of the T1M single crosses. This finding can be explained by the smaller total number of females than males, which increases the number of times each female is tested in single-cross combinations. Nevertheless, the mean of T1 single-cross prediction accuracies were 79, 90 and 83 percent of the T2 single-cross prediction for GY, PH, and SG, respectively. The mean accuracies of T0 single-cross prediction were 53, 79 and 65 percent of the mean accuracies of T2 single-cross prediction for GY, PH and SG, respectively. This indicates that performance of single crosses having at least one tested parent can be effectively predicted using genomic estimated GCA and SCA effects, but prediction accuracies suffer considerably if neither of the parents of a single cross are tested. This issue should be studied using larger population sizes – both in terms of more inter-connected bi-parental populations and progenies per population – to determine if population size can overcome parent representation in the training set.

Comparison of Single-Cross Prediction Methods

The published studies on prediction of single-cross performance have used covariance among tested and untested single crosses (Method 2a & 2b) to predict the performance of untested single crosses (Massman, *et al.* 2013a; Technow *et al.* 2014). In an alternative approach, we used genomic estimated GCA and SCA (Method 1a & 1b) to predict the performance of untested single crosses. Although the same information is input into the two different types of methods (i.e., additive genomic relationship matrices

of parents, dominance relationship matrix of the single crosses), the methods, however, differ in their underlying assumptions and in the way in which resulting predictions are calculated. Methods 1a and 1b use the three genomic relationship matrices separately in order to model the covariances of GCA effects of females, males and SCA effects of single crosses. Methods 2a and 2b, on the other hand, combine the covariance matrices to estimate the single-cross covariance through summation of the covariance between the female parents and the covariance between the male parents. The single-cross covariance derived in this way is valid under the assumption that allele frequencies and variance components between male and female populations are similar. These assumptions may not hold if male and female populations are separated for long period of time as in the case of heterotic groups. The comparison of prediction accuracies, however, showed that two groups of methods achieved similar prediction accuracies. In Method 1 of VanRaden, the male and female genomic relationships are weighted by the frequency of common reference allele, specifically minor allele, in the corresponding populations. This overcome the confounding effect of allele frequency differences in estimating single-cross covariance. The possible confounding effect of differences in variance components appears to be smaller if genomic relationships are weighted by the frequency of common reference allele.

The prediction accuracies for T2, T1, and T0 single crosses reported by Technow *et al.* (2014) and T2 and T1 single crosses reported by (Massman, *et al.* 2013a) are higher than corresponding accuracies observed in the present study. This discrepancy is likely due to the differences in population and family structure between the present study and

those previously reported. Consider two prediction scenarios, one from the previous studies and the one in our study. (Massman, *et al.* 2013a) and Technow *et al.* (2014) used single crosses made among a diverse set of established inbred parents. These inbred parents are likely to belong to distinct groups based on their single-cross performances. As Windhausen *et al.* (2012) reported, the prediction accuracy under such scenario results mostly from differences in mean performances between groups and less from genetic relationships between training and validation sets because a large amount of variation happens to be between groups compared to within groups. In our case, larger genetic variation was within families because there were many inbred progenies from each biparental family and there was one grandparent common between each pair of the families. Therefore, single-cross prediction accuracy mostly resulted from genetic relationships and less from differences among groups. However, the closer genetic relationship between the training and validation sets generated due to the common grandparent made it challenging to distinguish single-cross performances among closely related inbred progenies. In addition, the average number of single-cross combinations per parental line were higher in these studies which significantly increases the single-cross prediction accuracy as evidenced from higher prediction accuracy of T1F single crosses compared to T1M single crosses.

The Benefit of Modeling SCA

We observed similar prediction accuracies when including both GCA and SCA compared to predicting based on only GCA. Previous simulation study in maize have

reported increase in the prediction accuracy when including dominance in addition to additive marker effects (Technow *et al.* 2012). Similarly, in animals, Su *et al.* (2012) and Sun *et al.* (2014) showed that inclusion of nonadditive effects improved the predictions. The reported increase in prediction accuracies, however, were very small in spite the larger training population sizes used in these study. Significant increase in prediction accuracy including nonadditive effects was observed in combination of high proportion of dominance variance, larger training population size and closer genetic relationship between training and validation population (Denis and Bouvet 2013). The possibility of existence of these conditions, specifically first two, is rare in the breeding populations commonly used. This indicate that including nonaddtive effects may not benefit the accuracy of predictions. It is, however, important to note that all the above mentioned studies have used GBLUP model to investigate the importance of including nonadditive genetic effects for genomic prediction accuracy. GBLUP is a parametric models which requires assumption such as linearity and additivity of marker effects. These assumptions often do not hold in a typical breeding population limiting the ability of these models to precisely capture nonadditive genetic effects (Gianola *et al.* 2006). It would, therefore, be desirable to explore new methods e.g. nonparametric GS methods for utilizing nonaddtive genetic effects for genomic prediction of single-cross performance.

Prospects for Early-Stage Single-Cross Prediction

Overall, this study indicates that breeders should consider redesigning hybrid breeding programs to take advantage of genomic prediction. The early stages of maize

hybrid development consist of the generation of RILs or DHs from biparental families on each side of a heterotic pattern, followed by evaluation of their potential to serve as parents of hybrids. Traditionally, initial selections are conducted on the basis of topcross tests. Single crosses among selected inbred progenies are evaluated in later stages in the breeding pipeline. While this method has many advantages, one major disadvantage is that not all potential single crosses among breeding lines can be evaluated. Moreover, the addition of multiple years of topcross testing increases the time to hybrid release. The use of genomic prediction to identify superior single crosses could both shorten the length of time to hybrid release, and prevent the discarding of superior single crosses that just never happened to be phenotypically evaluated in the topcross system. We believe this can be achieved given the high prediction accuracies observed when both parents (T2) are included in the training set. Additionally, opportunity exists to optimize the genomic prediction of early-stage single crosses. Pedigree selection and frequent use of successful parents creates a family structure within typical hybrid maize breeding programs consisting of inter-connected biparental families. The results from this study demonstrate that single-cross genomic prediction methods even hold potential for separating single crosses from a common family background (Figure 6). The prediction accuracy for novel single-cross families was moderate to high and the addition of single crosses from the same family to the training set only minimally improved accuracy. Further study of the optimization of larger training sets through leveraging family structure could further improve the accuracy of genomic prediction of single cross.

Table 3. Family designations of nine single-cross families and number of single crosses belonging to each of the nine families. Biparental families are listed in the row and column headings. The numbers in the parentheses indicate numbers of recombinant inbred lines (RILs) or doubled haploid lines (DHLs) in the biparental family or number of single crosses in each single-cross family. Total number of single crosses per bi-parental family are displayed in the table margins.

	PHG47xPHG84 (35)	LH82xPHG47 (69)	LH82xPHG84 (67)	Total
PHJ40xPHG39 (8)	f1 (27)	f2 (39)	f3 (33)	**99**
B73xPHG39 (36)	f4 (51)	f5 (49)	f6 (49)	**149**
PHJ40xB73 (2)	f7 (21)	f8 (19)	f9 (24)	**64**
Total	**99**	**107**	**106**	**312**

Table 4. Mean, range, genetic variance and broad-sense heritability estimates in whole population as well as individual single-cross families for grain yield (GY; Mt/ha), plant height (PH; cm), and staygreen (SG; 1-10 rating)

Trait	Statistic	Single-cross Populations						
		Whole	f1	f2	f3	f4	f5	f6
GY	Mean	8.67	8.6	8.85	8.87	8.88	9.03	9.13
	Range	7.14-10.2	7.13-9.91	6.94-9.94	7.79-9.99	6.74-10.7	7.52-10.4	7.46-10.5
	$\sigma_g^2 \pm SE$	0.50 ± 0.07	0.9 ± 0.31	0.48 ± 0.18	0.25 ± 0.12	0.55 ± 0.19	0.51 ± 0.15	0.51 ± 0.16
	$H^2 \pm SE$	0.58 ± 0.04	0.80 ± 0.07	0.66 ± 0.09	0.53 ± 0.13	0.57 ± 0.10	0.71 ± 0.07	0.71 ± 0.07
PH	Mean	210.1	213.4	206.6	205.7	221.2	208.9	216.1
	Range	191 - 231	197 - 227	187-222	187-222	202-243	182-230	191 - 241
	$\sigma_g^2 \pm SE$	1.18 ± 0.1	0.71 ± 0.23	0.8 ± 0.22	0.95 ± 0.27	0.87 ± 0.21	0.9 ± 0.20	1.07 ± 0.24
	$H^2 \pm SE$	0.89 ± 0.01	0.88 ± 0.04	0.86 ± 0.04	0.90 ± 0.03	0.83 ± 0.04	0.90 ± 0.02	0.91 ± 0.02
SG	Mean	6.79	6.96	7.05	6.68	6.35	6.75	6.22
	Range	5.48-8.31	5.57-7.96	5.82-8.5	5.57-7.96	4.61-7.88	5.67-7.92	5.07-7.39
	$\sigma_g^2 \pm SE$	0.69 ± 0.07	0.36 ± 0.15	0.52 ± 0.16	0.26 ± 0.1	0.58 ± 0.14	0.38 ± 0.1	0.26 ± 0.07
	$H^2 \pm SE$	0.81 ± 0.02	0.67 ± 0.10	0.74 ± 0.07	0.68 ± 0.09	0.80 ± 0.04	0.78 ± 0.05	0.78 ± 0.05

Table 5. General combining ability variance of stiff stalk synthetic ($\sigma^2_{GCA\_F}$) and non-stiff stalk ($\sigma^2_{GCA\_M}$) inbred progenies and specific combining ability variance (σ^2_{SCA}) of single crosses between them.

Variance components	Grain yield	Plant height	Staygreen
$\sigma^2_{GCA\_F}$	0.22**	28.66**	0.12**
$\sigma^2_{GCA\_M}$	0.20**	34.48**	0.23**
σ^2_{SCA}	0.05**	2.6**	0.01**
$\sigma^2_{SCA}/(\sigma^2_{GCA\_F} + \sigma^2_{GCA\_M})$	0.12	0.04	0.03

** Significant at α = 0.001

Figure 3. Crossing scheme between RILs or DHLs derived from three biparental families representing the SSS (*y*-axis) and NSS (*x*-axis) heterotic groups. Colored boxes indicate the presence while unfilled boxes indicate absence of a particular single cross. Bold lines delineate single-cross families

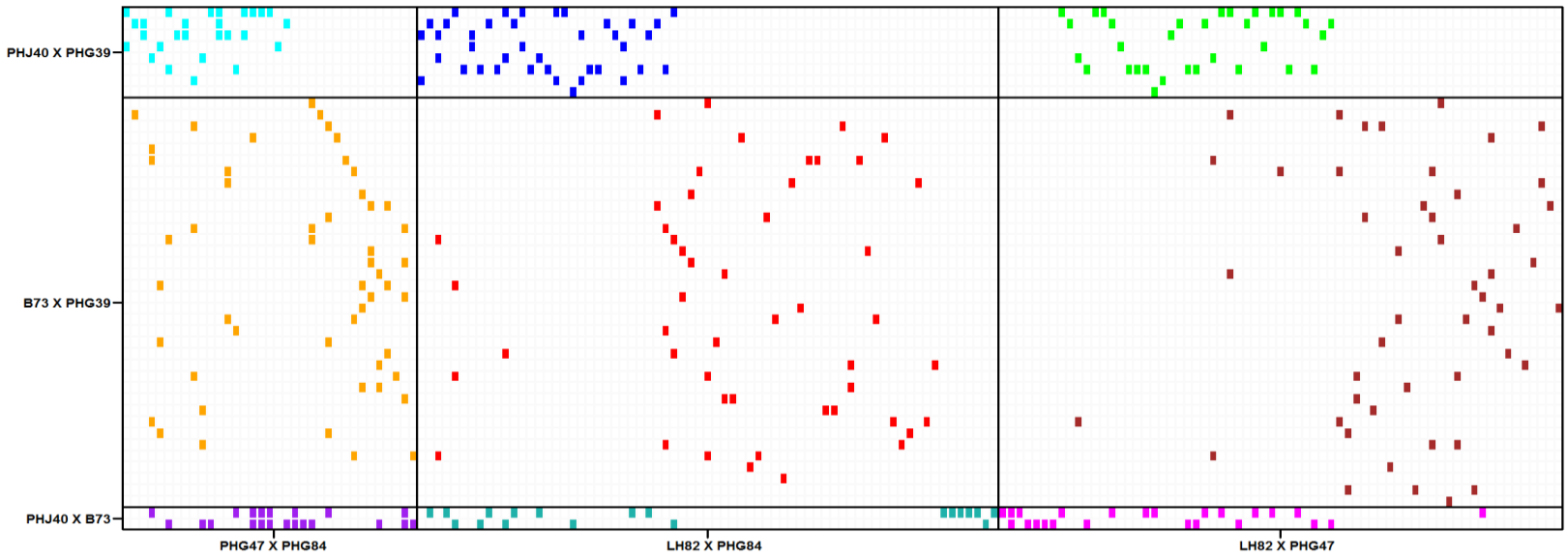


Figure 4. Schematic visualization of T2, T1F, T1M and T0 cross-validation scenarios. Each small square represents one single cross. Completely filled squares (T2) indicate that both male and female parents of a single cross contained in the validation set were tested, half-filled squares indicate either the female (T1F) or male parent (T1M) of single cross contained in the validation set was tested, and unfilled squares (T0) indicate that neither parent of a single cross contained in the validation set was tested.

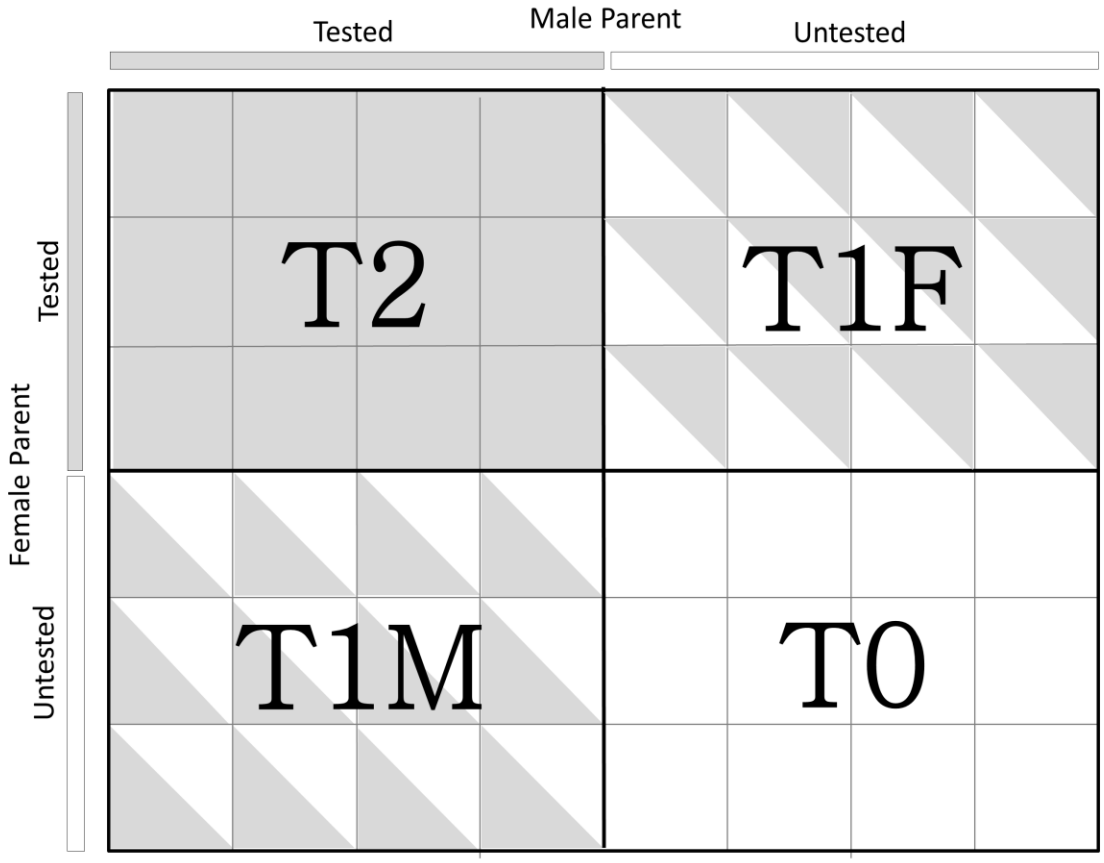


Figure 5. Prediction accuracy for T2, T1F, T1M and T0 cross-validation scenarios for traits grain yield (GY), plant height (PH) and staygreen (SG) obtained using the four methods 1a (Parent GCA), 1b (Parent GCA plus single-cross SCA), 2a (Additive genetic covariance among single crosses) and 2b (Additive plus dominance covariance among single crosses) as evaluated with training set of 250 and leave-one-individual-out cross-validation.

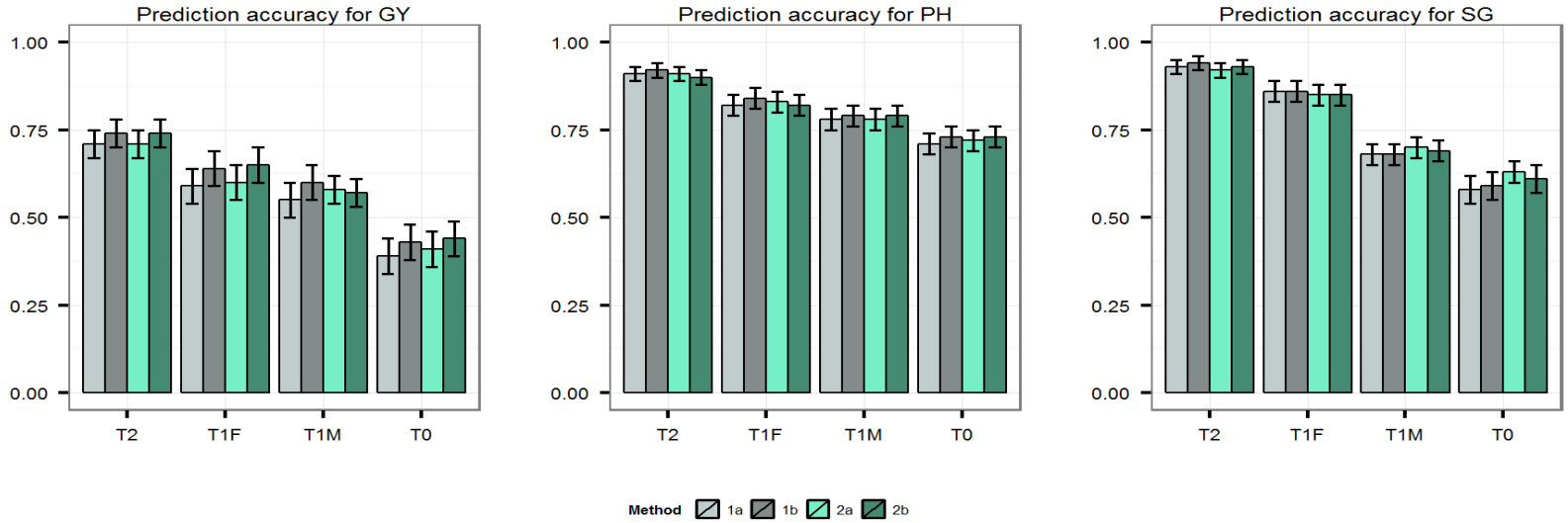


Figure 6. Mean prediction accuracy and standard errors of methods 1a (GCA) and 1b (GCA + SCA) in predicting performance of novel single-cross families. Two cross-validation schemes were used: leave-one-family out (bottom panel) and leave-one-individual out (top panel). Traits analyzed were grain yield (GY), plant height (PH), and stay green (SG). Standard errors were estimated using the bootstrap method.

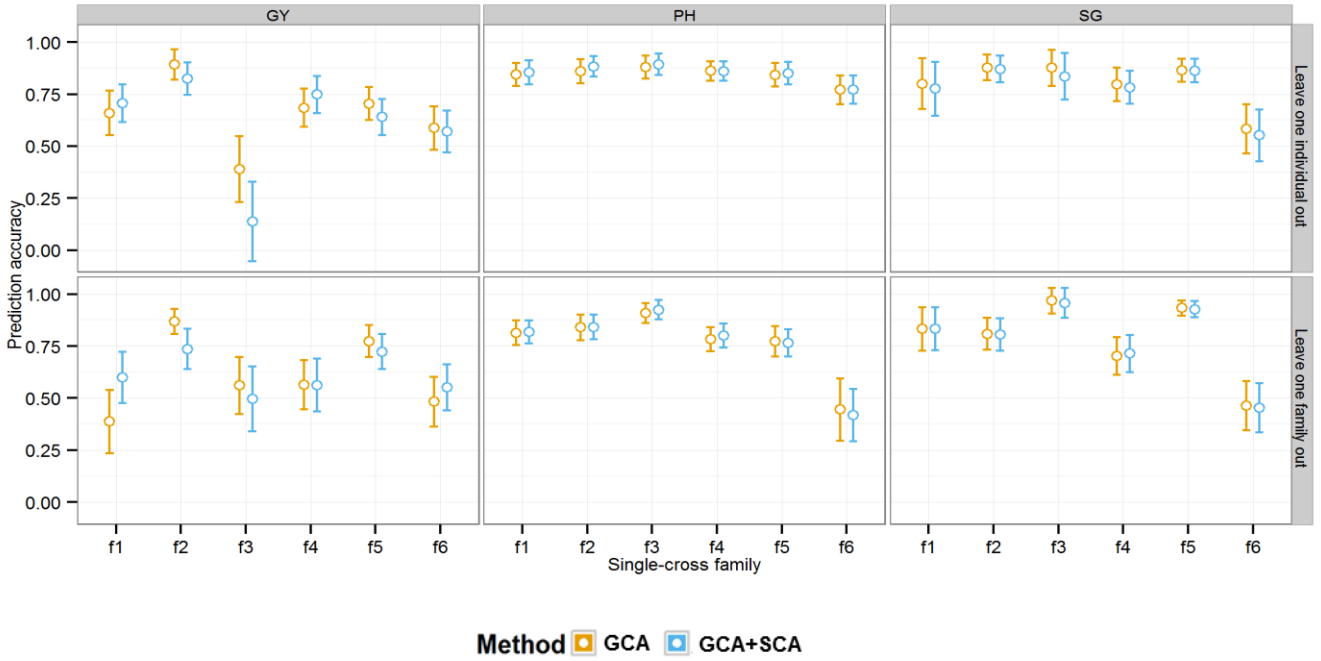
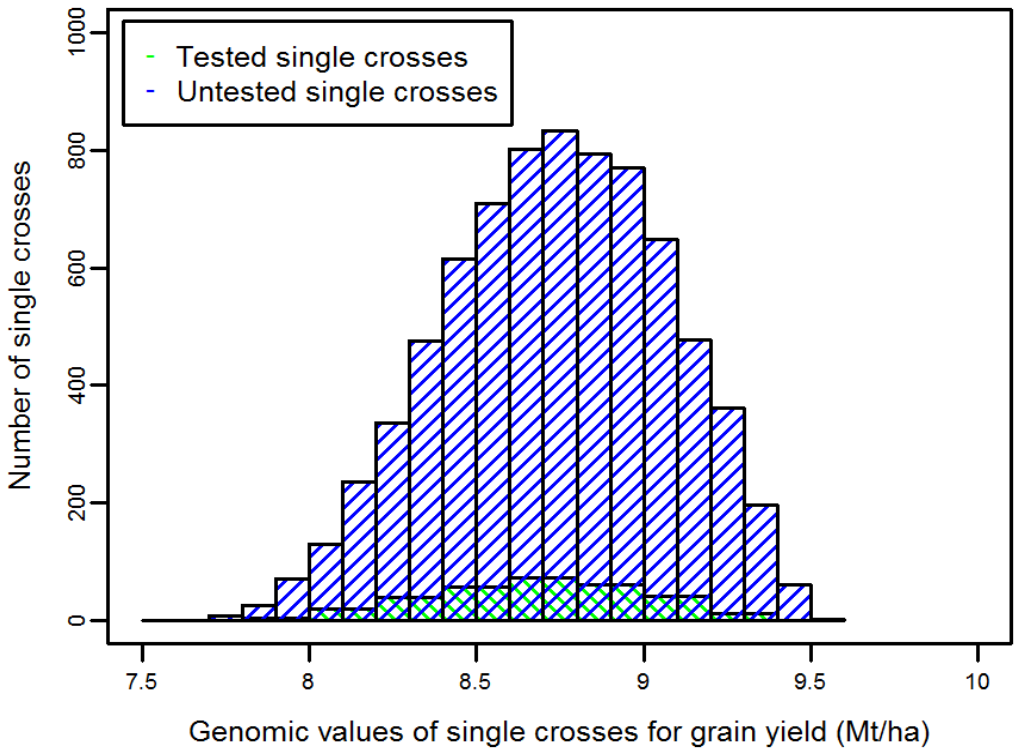


Figure 7. Distribution of genomic predictions for grain yield (GY) for all 7866 possible single crosses between the 46 SSS inbred progenies and 171 NSS inbred progenies.



Chapter 3: Evaluation of Nonparametric Models for Genomic Prediction of Early-Stage Single Crosses in Maize

Prediction of single-cross performance is extremely important because it is not feasible to evaluate all parental combinations in a hybrid breeding program. Recent simulation and experimental studies have shown great promise of genomic prediction of single-cross performance. These previous studies, however, were mainly focused on parametric genomic prediction models. In the present research, we investigated three nonparametric models: reproducing kernel Hilbert spaces (RKHS), support vector regression (SVR) and neural network (NN) in comparison to benchmark parametric model genomic best linear unbiased prediction (GBLUP) for predicting early-stage single crosses. Two separate datasets consisting of 481 and 312 single crosses generated between recombinant inbred lines/doubled haploid lines belonging to Iowa Stiff Stalk Synthetic (BSSS) and Non-Stiff Stalk Synthetic (NSSS) heterotic groups were used for this study. The genomic prediction models were trained to predict single-cross performance or general and specific combining abilities (GCA and SCA) of their parents. Genomic predictions were also compared to a mimicked topcross test and phenotypic estimates of GCA. Genome-based prediction of single-cross performance provided the highest predictive abilities followed by phenotype and topcross-based prediction. Predictive abilities of parametric and nonparametric models were nearly identical. All genomic prediction models showed good ability to predict GCA effects but could not predict SCA effects well. Our results, therefore, suggest that nonparametric models may

not provide additional advantage over parametric models for prediction of single-cross performance.

INTRODUCTION

The concepts of general combining ability (GCA) and specific combining ability (SCA), developed by Sprague and Tatum (1942), greatly enhanced the design of hybrid breeding programs (Hallauer and Miranda 1988; Fu *et al.* 2014). GCA describes the average performance of an inbred line when crossed to many other inbred lines, while SCA describes the deviation in performance of specific parental combinations from what is expected based on parental GCA. Generally, variation in GCA is due to additive genetic variation, while variation in SCA is due to dominance genetic variation, while additive-by-additive epistasis contributes to both. Although research has shown variation in GCA is the predominant source of variation among maize hybrids, variance in SCA is still important, especially for grain yield (Hallauer and Miranda 1988; Bernardo 1996a; Technow *et al.* 2014). This variation in SCA necessitates the evaluation of specific parental combinations (single crosses) to identify promising hybrids. Single-cross evaluation would ideally be performed using all possible crosses among parental lines. However, the sheer number of possible crosses – $N_1 \; x \; N_2$ for a factorial or $N(N-1)/2$ for a partial diallel – is far too large to evaluate in the early stages of a breeding program as the number of candidate parental lines is large. Various approaches have been proposed to evaluate the performance of candidate parents to be used in single crosses, including inbred *per se* performance, crossing to a tester line (topcross test) (Jenkins and

Brunson 1932), genetic distance estimate with molecular markers (Melchinger 1999), best linear unbiased prediction using pedigree or marker data (Bernardo 1994, 1996a) and markers associated with hybrid performance (Vuylsteke *et al.* 2000). Although some of these approaches have been useful, newer methods are needed to further improve the effectiveness of hybrid prediction (Schrag *et al.* 2009).

Models based on whole-genome marker data, known as genomic prediction models, have been developed to predict genetic values for complex traits (Meuwissen *et al.* 2001). Prospects of genomic prediction for single-cross performance have been investigated recently in maize (Massman, *et al.* 2013a; Technow *et al.* 2014). Genomic prediction for single-cross performance involves the prediction of genetic values of all possible single-cross combinations based on the genotypic data of all inbreds and phenotypic data on the subset of single crosses between them. Massman *et al.* (2013a) used ridge regression best linear unbiased prediction (RRBLUP) and genomic best linear unbiased prediction (GBLUP) to predict the performance of untested single crosses. The mean prediction accuracies were 0.87 for grain yield, 0.90 for grain moisture, 0.69 for stalk lodging, and 0.84 for root lodging. In another study, Technow *et al.* (2014) used GBLUP and BayesB to predict the untested single crosses. The prediction accuracies were ranged from 0.75 to 0.92 for grain yield and from 0.59 to 0.95 for grain moisture. These results suggest great promise for genomic prediction of single crosses.

The parental inbred lines used in the aforementioned studies on genomic hybrid prediction can be described as being from the advanced stages of a maize hybrid breeding

pipeline. That is, they were either elite parental lines already used in commercial hybrid production, or they resulted from multiple stages of selection based on testcross performance. Very few, if any, of the inbred lines used in these studies were from the same breeding cross as would be the case during the early stages of a maize hybrid breeding pipeline (Bernardo, 2002). The full potential of genomic prediction in this context would be better assessed by predicting and testing all possible single crosses among many random inbred lines derived from breeding crosses before selection in order to capture as much of the variation in the genetic space of the breeding program as a whole. Kadam *et al.* (2016) used GBLUP to predict the performance of early-stage single crosses made among random recombinant inbred lines (RILs) and doubled haploid lines (DHLs) from three biparental families from each heterotic group. Observed prediction accuracies for grain yield ranged from 0.39 to 0.77 depending on the model used and number of parents represented in the training set. The variance captured by modeling the SCA component was 12 % of the sum of the GCA variance. While this result indicates SCA is far from the predominant source of genetic variance in this type of population of single crosses, SCA variance still represents an important proportion of the variance. Further investigation into the potential of genomic prediction for single-cross performance with specific emphasis on prediction of SCA would be desirable.

Parametric genomic prediction models are not well suited to capture nonaddtive genetic effects (Gianola *et al.* 2006; Howard *et al.* 2014). These models regress the phenotype (y) on marker covariates (x_{ij}) using some type of regularization or variable selection procedure (de los Campos *et al.* 2013). Although both additive and nonadditive

effects can be included in these models by adding appropriate interactions between marker covariates, the partitioning of genetic value into additive, dominance and epistasis used in parametric models holds only under idealized conditions. These conditions often do not hold in a typical breeding program, limiting the effectiveness of these models to precisely capture nonadditve effects (Gianola *et al.* 2006). Moreover, the sheer dimensionality of the marker data used in genomic prediction could easily result in hundreds of millions or billions of interaction effects to estimate, which is a challenge not easily met by parametric methods (Gianola *et al*., 2010).

In an alternative approach, nonparametric models have been proposed to exploit nonaddtive genetic effects in genomic prediction (Gianola *et al.* 2006, 2010). These models are free of assumptions about the form of relationship between markers and phenotype, focus on prediction and seek a form that best fits the training data while maintain some generality for new data. This is in contrast to parametric methods where the focus is on parameter estimation rather than prediction. Due to this distinctive feature nonparametric models are expected to enable accounting for nonadditive genetics effects without explicitly modelling them and thus enhance the ability to predict phenotypes for complex traits (Gianola *et al.* 2010). Common nonparametric models used in the context of genomic prediction include reproducing kernel Hilbert spaces (RKHS) (Gianola *et al.* 2006, 2010), support vector regression (SVR) (Long *et al.* 2011) and neural network (NN) (Gianola *et al.* 2011). Empirical studies using these models have reported similar or greater accuracy compared to the benchmark models RRBLUP and GBLUP (Heslot *et al.* 2012; Pérez-Rodríguez *et al.* 2012; Crossa *et al.* 2013). Crossa *et al.* (2013) obtained a 5

– 18% improvement in prediction accuracies using RKHS as compared to GBLUP for grain yield in a testcross population of maize lines. Heslot *et al.* (2012) compared different parametric and nonparametric models for predicting various quantitative traits over eight different datasets. In their study, predictive ability of RKHS was greater than RRBLUP in 16 of the 18 comparisons made. In a study by Pérez-Rodríguez *et al.* (2012) mean prediction accuracies of RKHS and NN were better than Bayesian ridge regression for predicting days to heading and grain yield in a wheat dataset consisting of elite lines.

The suggested potential of nonparametric models to capture nonadditive effects makes them interesting candidates for genomic prediction of single-cross performance, which is considerably influenced by SCA (Massman *et al.* 2013a; Technow *et al.* 2014; Kadam *et al.* 2016). However, to our knowledge, no studies have yet compared these models for single-cross prediction. With these considerations, the main objective of the present study was to evaluate three nonparametric genomic prediction models -- RKHS, SVR and NN – for prediction of single crosses among random inbred lines from a limited number of families and compare them to GBLUP in terms of prediction accuracy. To address this objective, genomic prediction of single crosses was first compared to phenotype and topcross-based predictions to establish a baseline. Later, the four genomic prediction models were investigated in detail for single-cross prediction. Two different datasets of 481 and 312 single crosses made by randomly crossing RILs belonging to BSSS and NSSS heterotic groups were used for this study. The performance of single cross was predicted based on observed performances of tested single crosses or combining abilities of their parents.

MATERIALS AND METHODS

Plant Materials and Field Experiments

Dataset I

The germplasm in dataset I consisted of 481 single crosses between 89 RILs derived from six biparental families belonging to the BSSS heterotic group and 103 RILs derived from six biparental families belonging to the NSSS heterotic group. The parents of biparental families were Plant Variety Protection expired (ex-PVP) lines (Table 6). The six biparental families from the BSSS heterotic group were created by crosses among eight ex-PVP parents, and the six biparental families from the NSSS heterotic group were created by crosses among six ex-PVP parents. The RILs were F5-derived. Parent information and number of RILs and single crosses per family is displayed in Table 6. The RILs were randomly selected. The number of single crosses made per single BSSS and NSSS RIL ranged from 1 – 14 and 1 – 8, respectively. The mean number of single crosses per BSSS RIL was 5.4, and for NSSS it was 4.7. Field trials to evaluate agronomic traits of single crosses were conducted in Mead and York, NE during 2014 and Havelock and York, NE during 2015 for a total of four distinct environments. Trials included 450 single crosses in 2014 and 467 single crosses in 2015, with 436 single crosses evaluated in both the years. Thus, the total number of single crosses evaluated in at least one year was 481. The experimental design was a randomized incomplete block with two replications at each environment. Plots consisted of two rows 4.46 m in length and 0.76 m apart planted to a density of 88506 seeds per hectare. Plant height (PH) was

measured from base of the plant to collar of flag leaf at post anthesis. Three plants were randomly chosen per row and the mean of the six PH measurements was taken to represent the plot PH. Plots were machine harvested to determine grain weight per plot, which was converted to Mt ha^{-1} on a 155 g kg^{-1} moisture basis. Grain yield data points from plots having more than 10 lodged plants were discarded.

Dataset II

The germplasm in dataset II consisted of 312 single crosses made using an incomplete factorial design between 46 RILs or DHLs belonging to the BSSS heterotic group and 171 RILs or DHLs belonging to the NSSS heterotic group. The RILs or DHLs were derived from three interconnected biparental families per heterotic group. The RILs were in F6 selfing generation. Parent information and number of RILs/DHLs and single crosses per family is displayed in Table 7. The mean number of single crosses per individual BSSS inbred line was 6.9, whereas it was 1.8 for the NSSS inbred line. Single crosses comprising dataset II were evaluated in Illinois in 2012 at two locations, and in 2013 at three locations. Two locations were common between the years. The five location–year combinations were defined as separate environments. The experimental design was an $\alpha(0, 1)$-incomplete block design (Patterson and Williams 1976) with three replications at each environment. The phenotypic data were recorded for several agronomic traits. For this study, data on GY and PH were used for downstream analyses. GY was converted to Mt ha^{-1} on a 155 g kg^{-1} moisture basis. PH was measured post anthesis on a single representative plant determined by visually surveying the entire plot

before measurement. This dataset was used previously in Kadam *et al.* (2016) from which more detailed information on experimental design and measurement can be found.

Genotyping by Sequencing

Genotyping of RILs included as part of Dataset I was performed using genotyping-by-sequencing (Elshire *et al.* 2011). Briefly, five seeds of each RIL were planted in the greenhouse for leaf sample collection and pooled leaf tissue was immediately frozen in liquid nitrogen. DNA was extracted from lyophilized leaf samples using the Qiagen DNeasy Plant 96 kit. Library preparation and sequencing were performed at the Institute for Genomic Diversity (IGD) at Cornell University as described by Elshire *et al.* (2011). Single nucleotide polymorphisms (SNPs) were called from the raw sequence data using the TASSEL GBS Pipeline version 3.0 (Glaubitz *et al.* 2014) on the combined set of BSSS and NSSS RILs. SNPs with greater than 20% missing values and less than 5% minor-allele frequency were removed wherein heterozygous SNPs calls were set to missing values. Missing values were subsequently imputed using naïve imputation. Only SNPs that were polymorphic in both the BSSS and NSSS sets were retained, leaving 23,923 SNPs for statistical analysis.

The procedure used for DNA extraction, genotyping and SNP calling in dataset II was described in Kadam *et al.* 2016. The marker profiles of single crosses were inferred from their parental SNP information. SNPs were filtered for 5% minor-allele frequency. Only SNPs that were common to all three sets (BSSS RIL/DHs, NSSS RIL/DHs, and

single crosses were retained for further analysis. A total of 2273 SNPs remained after filtering.

Phenotypic Data Analysis

Analysis of phenotypic data across the environments was performed for each dataset using the following statistical model

$$y_{iklq} = \mu + g_i + (ge)_{ik} + r_{l(k)} + b_{q(kl)} + \varepsilon_{iklq} \quad ...(1)$$

where y_{iklq} is the phenotypic observation of the i^{th} single cross evaluated in the k^{th} environment in the l^{th} complete block (i.e. replication) and q^{th} incomplete block; μ is the grand mean; g_i is the effect of the i^{th} single cross; e_k represents the effect of the k^{th} environment; $(ge)_{ik}$ represents the interaction effect between the i^{th} single cross and k^{th} environment; $r_{l(k)}$ represents the effect of the l^{th} complete block nested within the k^{th} environment; $b_{q(kl)}$ represents the effect of the q^{th} incomplete block nested within the l^{th} complete block in the k^{th} environment; and ε_{iklq} represents the residual. Environment and replication nested within environment effects were modeled as fixed effects. All other effects were treated as random. The distributions of g_i and $(ge)_{ik}$ were assumed as follow: $g_i \sim N(0, I\sigma_g^2)$ and $(ge)_{ik} \sim N(0, I\sigma_{g\times e}^2)$. Error and incomplete block variances were allowed to be heterogeneous among environments. Stand count (SC) data was included as a covariate in model (1) for analysis of GY in dataset I. Inclusion of SC as a covariate reduced the error variance by four percent.

The restricted maximum likelihood estimates (REML) of all variance components were obtained using ASReml-R software (Butler *et al.* 2009). Significance of the

variance components was determined using likelihood ratio tests at 0.001 level of significance. The entry-mean heritability of each trait was estimated as: $H^2 = \sigma_g^2 / \left(\sigma_g^2 + \frac{\sigma_{g \times e}^2}{e} + \frac{\sigma_e^2}{re} \right)$ where, $\sigma_{g \times e}^2$ represents the variance of single cross × environment interaction effect, e is the number of environments, and r is the number of replications in each environment. The phenotypic data were unbalanced due to missing observations in dataset II. Therefore, e and re were substituted by the harmonic mean of number of observations per single cross within an environment, and harmonic mean of total number of observations per single cross as suggested by Holland *et al.* (2003).

BLUPs of single crosses obtained from equation (1) were used as observed single-cross performance for building genomic prediction models and evaluation of predictive ability as described in the next section. GCA and SCA effects were estimated using the following model:

$$y_{ijklq} = \mu + f_i + m_j + s_{ij} + e_k + (fe)_{ik} + (me)_{jk} + (se)_{ijk} + r_{l(k)} + b_{q(kl)} + \varepsilon_{ijklq} \quad \ldots(2)$$

where y_{ijklq} is the phenotypic observation for single cross between i^{th} female (BSSS line) and j^{th} male (NSSS line) evaluated in the k^{th} environment in the l^{th} complete block (i.e. replicate) and q^{th} incomplete block; μ is the grand mean; f_i is the GCA effect of i^{th} female; m_j is the GCA effect of j^{th} male; s_{ij} is SCA effect of corresponding single cross between the i^{th} female and j^{th} male; e_k represents the effect of the k^{th} environment; and $(fe)_{ik}$, $(me)_{jk}$ and $(se)_{ijk}$ represents the interaction effects of GCA of female, male and SCA of cross with k^{th} environment. The distributions of f_i, m_j and s_{ij} were assumed as follows: $f_i \sim N(0, I\sigma_{GCA\_f}^2)$, $m_j \sim N(0, I\sigma_{GCA\_m}^2)$ and $s_{ij} \sim N(0, I\sigma_{SCA}^2)$ where $\sigma_{GCA\_f}^2$ is

the variance of GCA effects of females, $\sigma^2_{GCA\_m}$ is the variance of GCA effects of males and σ^2_{SCA} is the variance of SCA effects. All remaining terms are as described in equation (1).

BLUPs of GCA and SCA effects obtained using equation (2) were treated as phenotype-based GCA and SCA estimates. These were used for building combining ability-based genomic prediction models as described in the following section.

Genomic Prediction Models

Three nonparametric GS models -- RKHS, SVR and NN -- were evaluated for prediction single-cross performance in comparison with GBLUP. The models were constructed so as to predict the single-cross performance or combining abilities of the single-cross parents. In the latter case, single-cross performance was derived by the summation of predicted GCA and SCA effects. The definitions of GCA and SCA are analogous to the main and interaction effects of parents in a standard linear model. As a result, they can be estimated directly by modelling single-cross performance as response variable, GCA as a main effect, and SCA as an interaction effect in linear mixed model. However, as nonparametric models do not impose a linear and additive relationship between the response and explanatory variables, GCA and SCA effects cannot be estimated when these models are fit to single-cross performance. Therefore, we used GCA and SCA BLUPs obtained using equation (2) as the response variable in building nonparametric models for combining ability-based single-cross prediction. To further

clarify, the basic structure of the nonparametric models for single-cross performance and combining ability based prediction is given follow.

1. *Single-cross performance based model*

$$y_{SC} = f(X_h) + \varepsilon$$

where, y_{SC} is a vector of BLUPs of single crosses obtained from equation (1) and $f(X_h)$ is a certain function genotype matrix (X_h) of single crosses inferred from parent genotype data and ε is a vector of residuals.

2. *Combining ability based model*

$$y_{CA} = f(X_p) + \varepsilon$$

where y_{CA} is vector of combining abilities. For GCA prediction, y_{CA} is a vector of female or male GCA effect estimates from model 2 and $f(X_p)$ is a certain function of genotype matrix (X_p) of females or males. For SCA prediction, y_{CA} is a vector of SCA effect estimates from model 2 and $f(X_p)$ is a certain function of heterozygote genotype matrix of single crosses. ε is as described above.

We first describe the GBLUP model and then describe the RKHS, SVR and NN models used in this study. In-depth descriptions of three nonparametric models can be obtained from the following references: RKHS (Gianola *et al.* 2006; De Los Campos *et al.* 2010); SVR (Long *et al.* 2011) and NN (Gianola *et al.* 2011).

Genomic best linear unbiased prediction

The GBLUP model with additive and dominance effects of single crosses has the form:

$$y = 1_n\mu + Z_1 u_A + Z_2 u_D + e \quad \text{... (3)}$$

where y is y_{SC}, 1_n is a n-length vector of ones, μ stands for the grand mean, Z_1 and Z_2 are the design matrices for the vectors of random additive and dominance genetic effects of n single crosses. The genetic effects have the co-variance structures $u_A \sim N(0, G_A\sigma_A^2)$ and $u_D \sim N(0, G_D\sigma_D^2)$, where σ_A^2 and σ_D^2 are the additive and dominance genetic variances, and G_A and G_D represents the $n \times n$ additive and dominance genomic relationship matrices, respectively. G_A was calculated following VanRaden (2008) as $G_A = \frac{WW'}{2\sum_{k=1}^{p} p_k(1-p_k)}$ where $W$is an $n \times m$ matrix with $w_{ij} = x_{ij} - 2p_j$ and x_{ij} is marker genotype of the i^{th} single cross for the j^{th} marker which is coded as 0, 1 or 2 for homozygous major allele, heterozygous, and homozygous minor allele states, respectively. G_D was calculated according to Da *et al.* (2014) as $G_D = \frac{HH'}{4\sum_{k=1}^{p} p_k^2(1-p_k)^2}$ where H is an $n \times m$ matrix with $h_{ij} = -2(1-p_j)^2$, if the individual is homozygous for major allele, $h_{ij} = 2p_j(1-p_j)$, if the individual is heterozygous and $h_{ij} = -2p_j^2$, if the individual is homozygote for minor allele. The mixed model equations to obtain the solutions of additive and dominance effects were as according to Zhao *et al.* (2013):

$$\begin{bmatrix} \hat{\mu} \\ \widehat{u_A} \\ \widehat{u_D} \end{bmatrix} = \begin{bmatrix} n & 1_n^T Z_1 & 1_n^T Z_2 \\ Z_1^T 1_n & Z_1^T Z_1 + G_A^{-1}\sigma_e^2/\sigma_A^2 & Z_1^T Z_2 \\ Z_2^T 1_n & Z_2^T Z_1 & Z_2^T Z_2 + G_D^{-1}\sigma_e^2/\sigma_D^2 \end{bmatrix}^{-1} \begin{bmatrix} 1_n^T y \\ Z_1^T y \\ Z_2^T y \end{bmatrix}$$

The summation of estimates of additive genetic effect ($\widehat{u_A}$) and dominance genetic effects ($\widehat{u_D}$) was considered as the predicted single-cross performance.

The GBLUP model for combining ability based prediction was as follows:

$$\mathrm{y} = 1_\mathrm{n}\mu + Z_f g_f + Z_m g_m + Zs + \mathrm{e} \quad \ldots (4)$$

where, y is y_{SC}, Z_f, Z_m and Z are the design matrices for the vectors of random GCA effects of females (i.e., g_f), males (i.e., g_m) and SCA effect (i.e., s). The covariance structures assumed for the random effects were $g_f \sim N(0, G_f\sigma_\mathrm{f}^2)$, $g_m \sim N(0, \mathrm{G}_m\sigma_\mathrm{m}^2)$ and $s \sim N(0, S\sigma_\mathrm{s}^2)$ where σ_f^2, σ_m^2 and σ_s^2 are the variances of GCA effects of females, GCA effects of males, and SCA effects, respectively. G_f, G_m and S are the additive genomic relationship matrices among females, males and dominance genomic relationship matrix among single crosses, respectively. G_f and G_m were calculated following VanRaden (2008) as described above. S was calculated as per Bernardo (2002). The mixed model equations to obtain g_f, g_m, and s were (Bernardo 2002):

$$\begin{bmatrix} \hat{\mu} \\ \widehat{g_f} \\ \widehat{g_m} \\ \hat{s} \end{bmatrix} = \begin{bmatrix} n & 1_\mathrm{n}^T\mathrm{Z_f} & 1_\mathrm{n}^T\mathrm{Z_m} & 1_\mathrm{n}^T Z \\ \mathrm{Z_f}^T 1_\mathrm{n} & \mathrm{Z_f}^T\mathrm{Z_f} + G_f^{-1}\sigma_e^2/\sigma_\mathrm{f}^2 & \mathrm{Z_f}^T\mathrm{Z_m} & \mathrm{Z_f}^T Z \\ \mathrm{Z_m}^T 1_\mathrm{n} & \mathrm{Z_m}^T\mathrm{Z_f} & \mathrm{Z_m}^T\mathrm{Z_m} + G_m^{-1}\sigma_e^2/\sigma_m^2 & \mathrm{Z_m}^T Z \\ Z^T 1_\mathrm{n} & Z^T\mathrm{Z_f} & Z^T\mathrm{Z_m} & Z^T Z + S^{-1}\sigma_e^2/\sigma_s^2 \end{bmatrix}^{-1} \begin{bmatrix} 1_\mathrm{n}^T\mathrm{y} \\ \mathrm{Z_f}^T\mathrm{y} \\ \mathrm{Z_m}^T\mathrm{y} \\ Z^T\mathrm{y} \end{bmatrix}$$

Both GBLUP models were implemented using ASReml-R software (Butler *et al.* 2009) to obtain REML of all variance components and solve the mixed linear model equations.

Reproducing kernel Hilbert space

Gianola *et al.* (2006) introduced the RKHS model for genomic prediction. In this model, the genomic relationship matrix used in GBLUP is replaced by a kernel matrix which enables non-linear regression in a higher-dimensional feature space. The RKHS model can be represented as follow:

$$y = 1_n\mu + K\alpha + \varepsilon$$

where K is a $n \times n$ reproducing kernel matrix whose entries are functions of marker genotypes of pairs of individuals. α is an $n \times 1$ vector of regression coefficients and ε is a vector of residuals of length n. The distributions of α and ε are $\alpha \sim N(0, K^{-1}\sigma_\alpha^2)$ and $\varepsilon \sim N(0, I\sigma_\varepsilon^2)$. The values of α are estimated by minimizing the objective function $(y - K\alpha)'(y - K\alpha) + \lambda\alpha' K\alpha$. The solution to minimizing the above function is (Morota and Gianola 2014):

$$\hat{\beta} = (K + hI)^{-1}y$$

We used a Gaussian kernel with $K_{ij} = \exp(-h\|x_i - x_j\|^2)$, where $\|x_i - x_j\|$ is the Euclidean distance between individuals i and j normalized between to 0 and 1, and h is a bandwidth parameter which controls the rate of decay of K_{ij} with increasing Euclidean distance.

The RKHS model was implemted using the R package BGLR (de los Campos and Pérez-Rodríguez 2013). The optimum value for h was chosen by performing cross validations using the whole dataset over a range of values from 10^{-6} to 10. The parameters *nIter* and *burnIn* specifying total number of iterations for Gibbs-sampler and

initial number of iterations to be discarded were set at 40,000 and 1000, respectively. All other parameters were specified according to the default settings.

Support vector regression

Support vector regression is a state of art machine learning algorithm for classification and regression problems. Maenhout *et al.* (2007), and Long *et al.* (2011) used SVR for genomic prediction in plant breeding, and Moser *et al.* (2009) for genomic prediction in animal breeding. SVR is a particular case of RKHS regression where the quadratic error loss function in RKHS is replaced by ε-insensitive loss function (González-Recio *et al.* 2014). The ε-insensitive loss function has the following form:

$$L(y_i - f(x_i)) = \begin{cases} 0, & \text{if } |y_i - f(x_i)| < \varepsilon \\ |y_i - f(x_i)| - \varepsilon, & \text{otherwise} \end{cases}$$

If the errors are less than ε, the loss function assigns zero loss. If the errors are larger than ε, the loss is equal to the difference between absolute error and ε. Thus, ε determines the number of support vectors used in the regression function. The objective function to be minimized with ε-insensitive loss is

$$C\sum_{i=1}^{n}(\xi_i - \xi_i^*) + \frac{1}{2}w'w$$

where C is a penalty parameter, $\xi_i \geq y_i - f(x_i) - \varepsilon$; $\xi_i^* \geq f(x_i) - y_i - \varepsilon$ and w is a vector of unknown weights i.e. regression coefficients. The minimizing solution to this objective function is given by $\hat{f}(X) = \sum_{i=1}^{n}(\alpha_i - \alpha_i^*)k(x, x_i)$, where k(.) is a kernel of choice and α_i and α_i^* are solutions to a nonlinear system of equations (Moser *et al.* 2009).

In our implementation of SVR, we used Gaussian radial basis kernel which has the form $k(x_i - x_j) = \exp(-\sigma\|x_i - x_j\|^2)$, where σ is the bandwidth parameter.

SVR was implemented using R package *kernlab* (Zeileis *et al.* 2004). The specified parameters were '*eps-svr*' (epsilon-regression) as the *type* and '*rbfdot'* (radial basis kernel "Gaussian") as the *kernel*. The optimal value of three tuning parameters required to solve the SVR regression i.e. the insensitivity zone (ε), penalty parameter (C) and bandwidth parameter (σ) were determined by grid search over a three dimensional parameter space. The grid values for ε ranged from 10^{-6} to 1; grid values of C and σ ranged from 10^{-6} to 10. All other parameters were set to the default values.

Neural network

In the field of statistical prediction a NN is a nonparametric prediction procedure based on how neurons in the brain work together to solve problems (Hastie *et al.* 2002). Gianola *et al.* (2011) first used NNs for genomic prediction. One of the basic and most commonly used forms of NNs for genomic prediction is the single hidden layer feed-forward NN. This form consists of an "input layer", a "hidden layer", and an "output layer". Predictions from this form of NN are obtained in two steps. In the first step, inputs are nonlinearly transformed in the hidden layer. This is accomplished by combining the inputs (x_{ij}) with weights (β^t) and an intercept (b_t) in each of *t* (t = 1, 2, 3,…*s*) neurons. This is followed by transformation of a linear score ($v_i^t = b_t + \sum_{j=1}^{p} x_{ij}\beta_j^t$) through a non-linear activation function, $z_i^t = g_t(v_i^t)$. In the second step, the response variable

(i.e., phenotype) is linearly regressed on the data derived predictors, $z_i^{[t]}$. The output function of NN can be represented as

$$y_i = \mu + \sum_{t=1}^{S} w_t z_i^t + \varepsilon_i = \mu + \sum_{t=1}^{S} w_t g_t \left(b_t + \sum_{j=1}^{p} x_{ij} \beta_j^t \right) + \varepsilon_i$$

where w_t is the weight of t^{th} neuron to the output, $g_t(.)$ is the activation function, and $\varepsilon_i \sim N(0, \sigma_e^2)$.

The NN model was implemented in this study using the R package *brnn* (Pérez-Rodriguez and Gianola 2013). A tangent hyperbolic activation function, $g_t(x) = \frac{\exp(2x)-1}{\exp(2x)+1}$, was used in this implementation. The model was fitted using a genomic relationship matrix rather than a SNP incidence matrix as the predictor (Gianola *et al.* 2011). The genomic relationship matrices were calculated as in GBLUP section. The number of epochs to train the model were set to be 30 except for SCA prediction where less than 10 epochs were used. More epochs generated singularities due to less variability in SCA covariance matrix. The optimal number of neurons was determined by cross-validation using the whole dataset. Neuron numbers ranged from 1 - 6 in the cross validation. All other parameters were set to the default values.

Cross-validations

First, predictive ability for genetic value of single crosses was assessed using phenotype, topcross and genome-based prediction. The phenotype-based predictions were calculated using parents GCA estimated from the phenotype data only. For topcross-

based prediction, we mimicked a single tester-based prediction (topcross test) commonly used in the early stages of a hybrid development pipeline. For this, performance of an untested single cross was predicted based on one randomly sampled tested single cross involving either male or female parent of the untested single cross. For each untested single cross, one tested single cross representing female parent and one tested single cross representing male parent were sampled to form two vectors of tested single crosses. These were later used separately to estimate the performance of untested single crosses. The mean correlation obtained was adjudged as the topcross-based single-cross predictive ability. Random sampling of a tested single cross for each parent was repeated 10 times. Genome-based single-cross predictions were based GBLUP model as in equation (4).

To further compare genome and phenotype-based prediction, four types of single crosses depending upon the tested or untested parents were distinguished for cross-validations. These single crosses were named as: T2 - both parents tested in other single-cross combinations; T1F – only female parent tested in other single-cross combinations; T1M - only male parent tested in other single-cross combinations; and, T0 - neither parent tested in other single-cross combinations. Also, the notation T1 was used to denote singles crosses having either male or female parent tested.

Leave-one-out cross-validations were performed to evaluate phenotype, topcross and genome-based predictive abilities of single crosses. For all cross validations, the predictive abilities were estimated as the Pearson's correlation coefficient between the observed and predicted single-cross performance. The standard errors of the predictive

abilities were estimated using the bootstrap method implemented in the R package, boot (Canty and Ripley 2012). The number of bootstrap samples used equal to 200.

RESULTS

Variance Components

Genetic variances among single crosses were significant for GY and PH in both datasets (Table 8). As expected, variance in GCA was far more important than variance in SCA, but SCA variance still accounted for 12-14 % of the total genetic variance for GY and 4-9 % for PH across datasets (Table 9).

Comparisons of Phenotype-, Topcross- and Genome-Based Predictive Abilities

We initially explored the single-cross predictive abilities based on the phenotype, topcross performance with single inbred tester (topcross-based prediction) and genome-based prediction obtained with GBLUP. A total of 192 (89 females and 171 males) and 217 (46 females and 171 males) single crosses are required for topcross-based prediction in dataset I and II, respectively. To compare the topcross-based prediction with phenotype- and genome-based prediction, we calculated the single-cross predictive abilities of the latter two approaches with training set sizes of 192 and 217 in dataset I and II, respectively, in order to eliminate confounding effects of population size. The highest predictive abilities were obtained with genome-based prediction followed by phenotype-based prediction, while lowest predictive abilities were obtained with topcross-based prediction (Table 10). For GY, the average improvement of genome-based prediction over topcross-based prediction was 97%, representing a huge

improvement in terms of identifying superior single crosses. The genome-based approach even beat the phenotype-based approach, amounting to a 9% improvement for GY on average across the two datasets.

Phenotype and Genome-Based Predictive Abilities for T2, T1 and T0 Single Crosses

We further investigated the phenotype and genome-based predictive abilities specifically for T2, T1 and T0 single crosses. The predictive abilities were considerably differed depending upon the number of tested parents of single crosses (Figure 8). The predictive abilities for T2 single crosses were highest followed by T1F, T1M and T0 for both traits in each datasets. The mean genome-based predictive abilities for T1 and T0 single crosses were 85 and 69 percent of mean predictive abilities for T2 single crosses for GY. Similarly for PH, the mean genome-based predictive abilities for T1 and T0 single crosses were 87 and 73 percent of mean predictive abilities for T2 single crosses.

The mean (averaged over T2 and T1 single crosses across two datasets) genome-based predictive abilities were 6 % and 12 % higher than the mean phenotype-based predictive abilities for GY and PH respectively. The advantage of genome-based prediction was higher especially when one of the parents of single crosses were untested. For T1 single crosses, the mean genome-based predictive abilities were 27 % and 32% better than phenotype-based predictive abilities for GY and PH respectively. Moreover, the mean genome-based predictive abilities for T0 single crosses were 0.45 for GY and 0.62 for PH which cannot be predicted based on phenotype data. The benefit of genome

based prediction over phenotype-based prediction was even more pronounced with genome-based methods providing highest predictive abilities.

Influence of Tuning Parameters

Ten-fold cross-validations in replicates of five were performed separately for GY and PH in each dataset to find the optimum values of tuning parameters for each nonparametric model. The sensitivity of these models to the corresponding tuning parameter values was explored. The models were robust over a considerable range of the parameter values. Based on a preliminary analysis, appropriately spaced grid values were selected and a cross-validation based grid search was performed on the whole data set. Predictive abilities of the three models were varied considerably over the range of tuning parameter values investigated (Figure 9). However, maximum predictive (± 0.01) ability was observed over a broad range of tuning parameters (RKHS and NN) or with many different combinations of tuning parameters (SVR). In the case of RKHS, predictive ability for both GY and PH in datasets I and II was maximized between *h* values 0.005 – 2 (Figure 9A). Similarly, in case of SVR, many different combinations of ε, C and σ values provided maximum predictive ability (Figure 9B). However, it appears that the combination of three parameter values providing maximum predictive abilities were specific to dataset and trait. The predictive ability of NN varied only slightly over different *s* values from 1 to 6, with a maximum difference of 0.04 across traits and datasets (Figure 9C). The optimum values of different tuning parameters for GY and PH in each dataset used for further analysis are provided in Table 11. We also compared the

predictive abilities of the three models with the optimum values of tuning parameters against the default values provided by the software packages used for implementing these models (Table 12). Optimum tuning parameters improved the predictive abilities of SVR and NN, but many times the improvement was marginal. The optimum and default *h* values were same (0.5) for RKHS resulting in the equal predictive abilities.

Comparison of GBLUP and Nonparametric Models

The single-cross predictive abilities of GBLUP and three nonparametric models differed slightly when the models were built on parents combining abilities (Figure 8). The predictive ability of GBLUP was highest followed by RKHS and SVR while NN provided lowest predictive abilities. The difference in predictive abilities between GBLUP and other three methods increased with increase in number of untested parents of single crosses. In case of GBLUP, including both GCA and SCA provided similar (± 0.01) predictive abilities. The predictive abilities of RKHS, SVR and NN lowered when including both GCA and SCA compared to predicting based on GCA only.

When the models were built on single-cross performance, the predictive abilities of GBLUP, RKHS and SVR were similar while NN provided lower predictive abilities (Figure 8). The predictive abilities of three nonparametric models built of single-cross performance were higher compared to building these models on parents combining abilities. The predictive abilities of GBLUP of combining ability were mostly similar to GBLUP of single-cross performance with little advantage for the former especially for the T1 and T0 single crosses and the trait PH.

DISCUSSION

In the early stages of hybrid breeding, RILs or DHLs are generated from several biparental families for testing in hybrid combinations. As the number of possible single crosses are too large to evaluate at this stage, initial selections of lines for hybrid performance are traditionally performed based on topcross test using a single inbred tester from the opposite heterotic group with evaluation of specific hybrid combinations occurring in the advanced stages (Hallauer and Miranda 1988). While the topcross test is easy compared to making and evaluating many pairwise crosses, the additional generations of topcross testing can increase the time required for commercial hybrid development. There is also a possibility of losing some unique potential single crosses due to discarding of lines in the early stages based only on topcross test data. Therefore, it would be desirable to evaluate the lines based on predicted single-cross performance in the early stages. In the present research, we initially evaluated three approaches for early-stage single-cross prediction: 1) phenotype-based prediction, 2) topcross-based prediction and 3) genome-based prediction. Phenotype- and genome-based predictive abilities were substantially higher than topcross-based prediction, indicating the importance of evaluating lines in single crosses. In a further comparison, genome-based prediction provided an advantage over phenotype-based prediction for T2, T1F, T1M single crosses. Moreover, the genome-based approach predicted T0 single crosses with moderate to high accuracy for which the phenotype-based approach has no ability to predict.

We next evaluated three nonparametric genomic selection models: RKHS, SVR and NN in comparison with GBLUP for prediction of early-stage single crosses. The predictive abilities of nonparametric models were remarkably similar to GBLUP. None of the nonparametric models provided advantage over GBLUP for single-cross prediction.

Selection of Tuning Parameters

For the implementation of nonparametric models, different tuning parameters needs to be provided. These tuning parameters include h for RKHS, ε, C and σ for SVR and s for NN. Generalized cross validation has been a broadly accepted method to determine the optimal values of these tuning parameters (Maenhout *et al.* 2007; Moser *et al.* 2009; Long *et al.* 2011; Heslot *et al.* 2012). We used the same approach to tune the *h* for RKHS, ε, C and σ for SVR and *s* for NN. The predictive abilities of these models were higher with optimum values chosen based on cross validation compared to default values (e.g. those used in the respective software packages for these models). This results is in accordance with previous studies which suggested optimal selection of tuning parameters to improve the prediction accuracy (Gianola and van Kaam 2008; Long *et al.* 2011). Additionally, we also observed that there is a great flexibility to select optimal values of tuning parameters as these models provided maximum predictive ability over a broad range of tuning parameters (RKHS and NN) or with many different combinations of tuning parameters (SVR).

Comparison of GBLUP and Nonparametric Models for Single-Cross Performance Prediction

Nonparametric genomic prediction models do not impose strong assumptions on genotype-phenotype relationships which is unlike parametric genomic prediction models that assume on priori a certain form relationship between genotype and phenotype (González-Recio *et al.* 2014). This is expected to enable nonparametric genomic prediction models to more effectively capture nonadditive effects, thus providing better predictions than parametric genomic prediction models (Gianola *et al.* 2006). In the present study, however, the predictive abilities of GBLUP, RKHS, and SVR were similar, and NN provided comparatively lower predictive abilities when the models were built on observed single-cross performance. Also, Maenhout *et al.* (2007) who previously investigated SVR for single-cross prediction found similar predictive abilities of SVR and GBLUP for grain yield, moisture content and days to flowering. Similar performances GBLUP and nonparametric models suggest that either nonadditive genetic effects are too small to capture, or that nonparametric models do no capture them well. In the present study, the proportion of SCA variances for GY and PH were smaller compared to GCA variances (Table 9). Similar estimates of GCA and SCA variances were reported among single crosses between dent and flint heterotic group inbreds (Technow *et al.* 2014) and BSSS and NSSS heterotic groups inbreds (Massman *et al.* 2013a). The low proportion of SCA variance can be expected for single crosses between genetically diverse heterotic groups (Reif *et al.* 2007).

On the other hand, genomic prediction studies in wheat have found an advantage in nonparametric models when compared to RRBLUP (Heslot *et al.* 2012; Pérez-Rodríguez *et al.* 2012) or GBLUP (Jiang and Reif 2015), suggesting that epistatic genetic variation is more important in wheat as compared to maize. One possible explanation for this could be that self-pollination has helped to maintain favorable epistatic gene combinations. Crossa *et al*., (2013) found better performance of RKHS than GBLUP for predicting testcross performance of 504 maize DHL crossed with an elite single-cross tester from opposite heterotic group. The differences in predictive abilities of RKHS and GBLUP were 0.08 for GY, 0.04 for days to anthesis and 0.03 for anthesis silking interval. They implemented RKHS model with kernel averaging approach (de Los Campos *et al.* 2010) while single kernel used in the present study. To test potential advantage of kernel averaging for single cross prediction, we performed single-cross prediction using RKHS model with kernel averaging approach as implemented in Crossa *et al*., (2013). The predictive abilities were similar or slightly lower compared to those obtained with single kernel chosen based on cross validations (results not shown). It therefore appears that use of DHL which involve limited generations of recombination and the use of single tester may have enabled to capture epistatic effects in their study.

Among the three nonparametric models, the predictive abilities of RKHS and SVR were remarkably similar while NN has lower predictive ability. Similar predictive abilities of RKHS and SVR can be explained by similar kernel definitions in both models. Specifically, these models use Gaussian kernel but different loss functions to solve optimization problem. The quadratic loss function is used in RKHS which is

replaced by ε-insensitive loss function in SVR. The similar predictive abilities of RKHS and SVR were also reported previously (Moser *et al.* 2009; Neves *et al.* 2012). The predictive ability of SVR was lower compared to RKHS in Heslot *et al.* (2012). One possible reason could be the use of linear kernel for SVR in their study while Gaussian radial basis kernel was used here. The low predictive ability of NN compared to other models may be due to a problem with overfitting. Unlike RKHS and SVR, which use single basis function defined on priori in regression, NN is based on combinations of many weak functions which are learned from the data (Gianola *et al.* 2011). Therefore, NN are more flexible and have the ability to learn any complex function from the data. For the same reason, however, it is also susceptible to overfitting, which decreases the predictive ability in the test data (Heslot *et al.* 2012).

GBLUP vs Nonparametric Methods for Parent's Combining Ability Based Prediction

When the nonparametric models were built on combining abilities of single-cross parents, their predictive abilities were lower compared to corresponding GBLUP predictive abilities. This difference in predictive abilities slightly increased from T2, T1F, T1M to T0 single-cross prediction scenarios (Figure 8). This could have resulted from different construction GBLUP and nonparametric models for combining ability based prediction rather than actual differences in abilities of these models to predict the genetic values. In case of GBLUP, the phenotype data on single crosses and marker derived covariance matrices among the parents of single crosses i.e. G_f, G_m and S were used

simultaneously to estimate the GCA and SCA of tested parents. In case of nonparametric models, GCA and SCA of tested parents were determined based on phenotype data on their single crosses alone. The genotypic data on tested parents and their GCA and SCA values were later used to train the model for predicting the GCA and SCA of untested parents. The use of marker derived covariance matrices in determining GCA and SCA can cause the resulting estimates to reflect more genetic component compared to determining them based only on phenotypic data where phenotype information solely rules the values of these estimates. Therefore, GCA and SCA estimated including genomic covariance among the parents can be more reliable for estimating the GCA and SCA of untested parents based on their marker information. The benefit of genomic estimated GCA and SCA increases when the parents are untested or tested in fewer single crosses. This could be because, when the parents are tested in fewer single crosses, the phenotype based GCA can be more biased by SCA compared to genotype-based GCA and SCA estimates. Such confounded GCA and SCA estimates are relatively less useful for estimating the GCA and SCA of untested parents.

The predictive abilities of nonparametric models were further decreased when SCA was included. The nonparametric models are expected to capture nonadditive effects by approximating a true form of relationship between marker and genotypic value. Hence, very low correlation between observed and predicted SCA from nonparametric models may be caused by inability to predict SCA separately (Table 13). Similar results were obtained when the tuning parameters were optimized specific to SCA (results not

shown). Overall, it appears that nonparametric models may not provide additional advantage over GBLUP in the context of single-cross prediction in maize.

GBLUP of Single-Cross Performance vs Combining Ability

Although GBLUP model built on combining abilities the parents or effects of single crosses have been considered as equivalent for single-cross prediction, there are subtle differences in underlying assumptions of these methods which can affect the predictive ability. GBLUP of single-cross performance assumes that marker effects are same in the male and female populations while GBLUP of combining ability assumes that marker effects are different between male and female populations (considering corresponding equivalent RRBLUP models). If the male and female populations are separated for a long period of time as in the case of heterotic groups in maize, the linkage phases between marker and QTL can change between populations. Hence, the assumption of equal marker effects may not hold. In the present study, however, GBLUP of single-cross performance and combining ability has provided similar predictive abilities with former giving little advantage for the PH (in case of traits) and T1 and T0 single crosses (in case of single cross types). These results are in accordance with a simulation study by Technow *et al.* (2012) who also reported only minor improvement in prediction accuracy when modelling marker effects specific to dent and flint populations.

Choice of Genomic Prediction Model for Single-Cross Performance

The optimal genomic prediction model is expected to have following characteristics 1. Highest possible accuracy, 2. Robust across traits and datasets, 3. Easy

to implement, and 4. Computationally efficient (Heslot *et al.* 2012). Based on these characteristics, the results of this study suggest that GBLUP built on combining abilities or genetic effects of single crosses is most appealing. None of the nonparametric models outperformed GBLUP. NN may not be recommended for single cross prediction because of low predictive ability compared to other models. Although RKHS and SVR have similar predictive ability as GBLUP, it could be computationally intensive to optimize the tuning parameters specific for trait and dataset.

Table 6. The parents of biparental families, number of RILs in each biparental family, number of single crosses for each of thirty six family wise cross combinations in dataset I. The total number of single crosses per biparental family are listed in the margins.

	PHG50 X LH123 (20)	PHZ51 X LH123 (14)	PHZ51 X PHG47 (14)	LH109 X LH123 (18)	LH109 X LH59 (20)	LH123 X LH59 (17)	**Total**
LH132 X B73 (16)	26	13	13	17	8	6	**83**
PHB47 X PHK29 (19)	19	13	14	15	16	13	**90**
PHB47 X PHW52 (17)	29	18	13	21	23	19	**123**
PHJ40 X PHW52 (3)	6	2	0	2	5	2	**17**
LH74 X LH132 (14)	8	3	9	11	8	8	**47**
PHG86 X PHW52 (20)	25	16	19	20	26	15	**121**
Total	**113**	**65**	**68**	**86**	**86**	**63**	**481**

Table 7. The parents of biparental families, number of RILs in each biparental family, number of single crosses for each of nine family-wise cross combinations in dataset II. The total number of single crosses per biparental family are listed in the margins.

	PHG47xPHG84 (35)	LH82xPHG47 (69)	LH82xPHG84 (67)	**Total**
PHJ40xPHG39 (8)	27	39	33	**99**
B73xPHG39 (36)	51	49	49	**149**
PHJ40xB73 (2)	21	19	24	**64**
Total	**99**	**107**	**106**	**312**

Table 8. Mean, range, variance components and heritabilities for grain yield (GY) and plant height (PH) in dataset I and dataset II.

Characteristics	Dataset I		Dataset II	
	Grain Yield (MT/ha)	Plant Height (cm)	Grain Yield (MT/ha)	Plant Height (cm)
Mean	10.97	243.48	8.67	210.1
Range	8.95 – 13.24	203.25 – 277.89	7.14 – 10.2	191 – 231
σ_g^2	0.71 ± 0.06	158.48 ± 10.8	0.50 ± 0.07	118.08 ± 10.6
H^2	0.76 ± 0.02	0.95 ± 0.004	0.58 ± 0.04	0.89 ± 0.01

Table 9. General combining ability variance of females (σ_f^2), males (σ_m^2) and specific combining ability variance (σ_s^2) of single crosses between them in dataset I and dataset II.

Variance components	Dataset I		Dataset II	
	Grain Yield (Mt/ha)	Plant Height (cm)	Grain Yield (Mt/ha)	Plant Height (cm)
σ_f^2	0.20 ± 0.05	50.16 ± 9.70	0.22 ± 0.07	28.66 ± 7.74
σ_m^2	0.09 ± 0.03	51.86 ± 9.15	0.20 ± 0.06	34.48 ± 5.92
σ_s^2	0.04 ± 0.01	9.53 ± 1.00	0.05 ± 0.01	2.6 ± 0.77
$\sigma_s^2/(\sigma_f^2 + \sigma_m^2)$	0.14	0.09	0.12	0.04

Table 10. Phenotype, topcross and genome based predictive abilities for grain yield (GY) and plant height (PH) in dataset I and II.

Dataset	Trait	Phenotype-based prediction		Topcross-based prediction		Genome-based prediction	
		Accuracy	SE	Accuracy	SE	Accuracy	SE
Dataset I	GY	0.66	0.02	0.36	0.04	0.71	0.03
	PH	0.71	0.02	0.38	0.04	0.76	0.02
Dataset II	GY	0.48	0.04	0.27	0.05	0.53	0.04
	PH	0.75	0.03	0.39	0.05	0.81	0.02

Table 11. Cross-validation based optimum values of tuning parameters of the three nonparametric methods RKHS, SVR and NN for grain yield (GY) and plant height (PH) in two datasets

Dataset	Trait	RKHS	SVR			NN
		h	C	ε	σ	s
I	GY	0.5	5	0.005	0.00005	4
	PH	0.5	10	0.05	0.000005	1
II	GY	0.5	5	0.05	0.0001	1
	PH	0.5	10	0.005	0.0001	4

Table 12. Predictive abilities of three nonparametric models obtained with optimum values of tuning parameters chosen by cross validations and default values provided by respective software packages used to implement these models

Dataset	Trait	RKHS		SVR		NN	
		Optimum	Default	Optimum	Default	Optimum	Default
I	GY	0.78	0.78	0.78	0.76	0.76	0.76
	PH	0.86	0.86	0.86	0.82	0.84	0.83
II	GY	0.56	0.56	0.58	0.51	0.52	0.49
	PH	0.82	0.82	0.83	0.79	0.81	0.80

Table 13. Correlation between observed and predicted GCA and SCA effects with GBLUP and three nonparametric methods RKHS, SVR and NN for grain yield (GY).

Hybrid Type	GBLUP		RKHS		SVR		NN	
	GCA	SCA	GCA	SCA	GCA	SCA	GCA	SCA
T2	0.931	0.125	0.896	-0.131	0.961	-0.070	0.916	-0.164
T1F	0.855	0.021	0.824	-0.022	0.865	-0.033	0.786	-0.137
T1M	0.822	0.154	0.776	-0.127	0.829	-0.075	0.781	-0.103
T0	0.737	0.048	0.704	-0.093	0.720	-0.033	0.656	-0.107

Figure 8. Single-cross predictive abilities based on phenotype (dashed lines) and four genome based methods GBLUP, RKHS, SVR and NN for T2, T1F, T1M and T0 single crosses for grain yield (GY) (A) and plant height (PH) (B) in dataset I and II as evaluated by leave-one-out cross-validation (LOOCV).

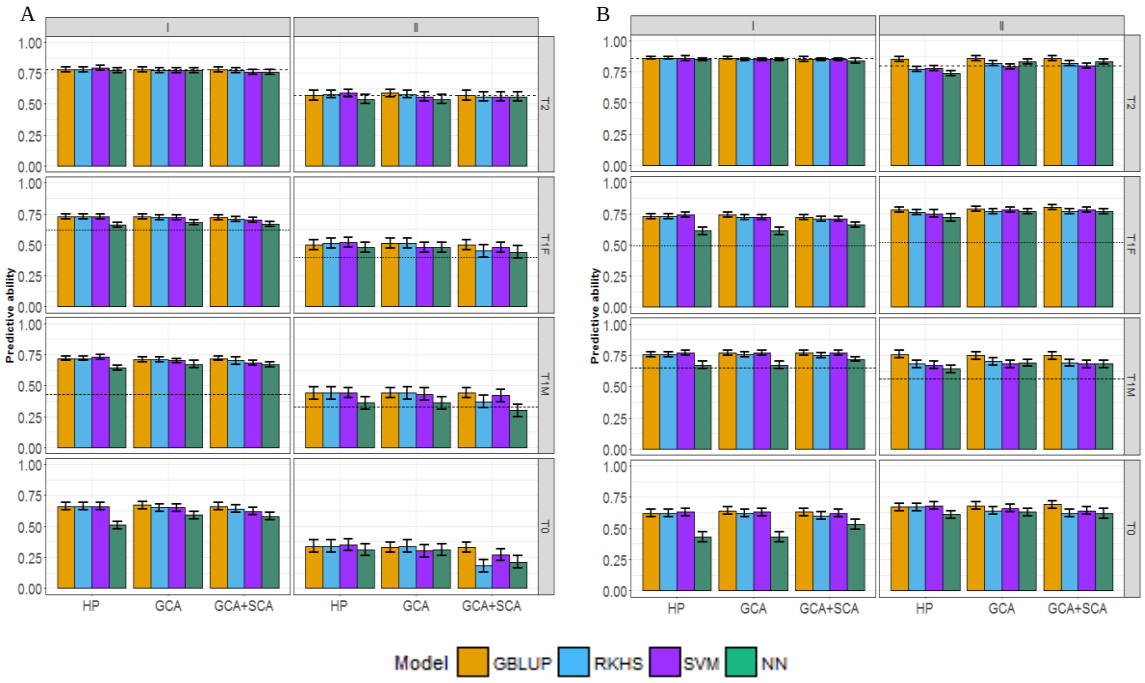
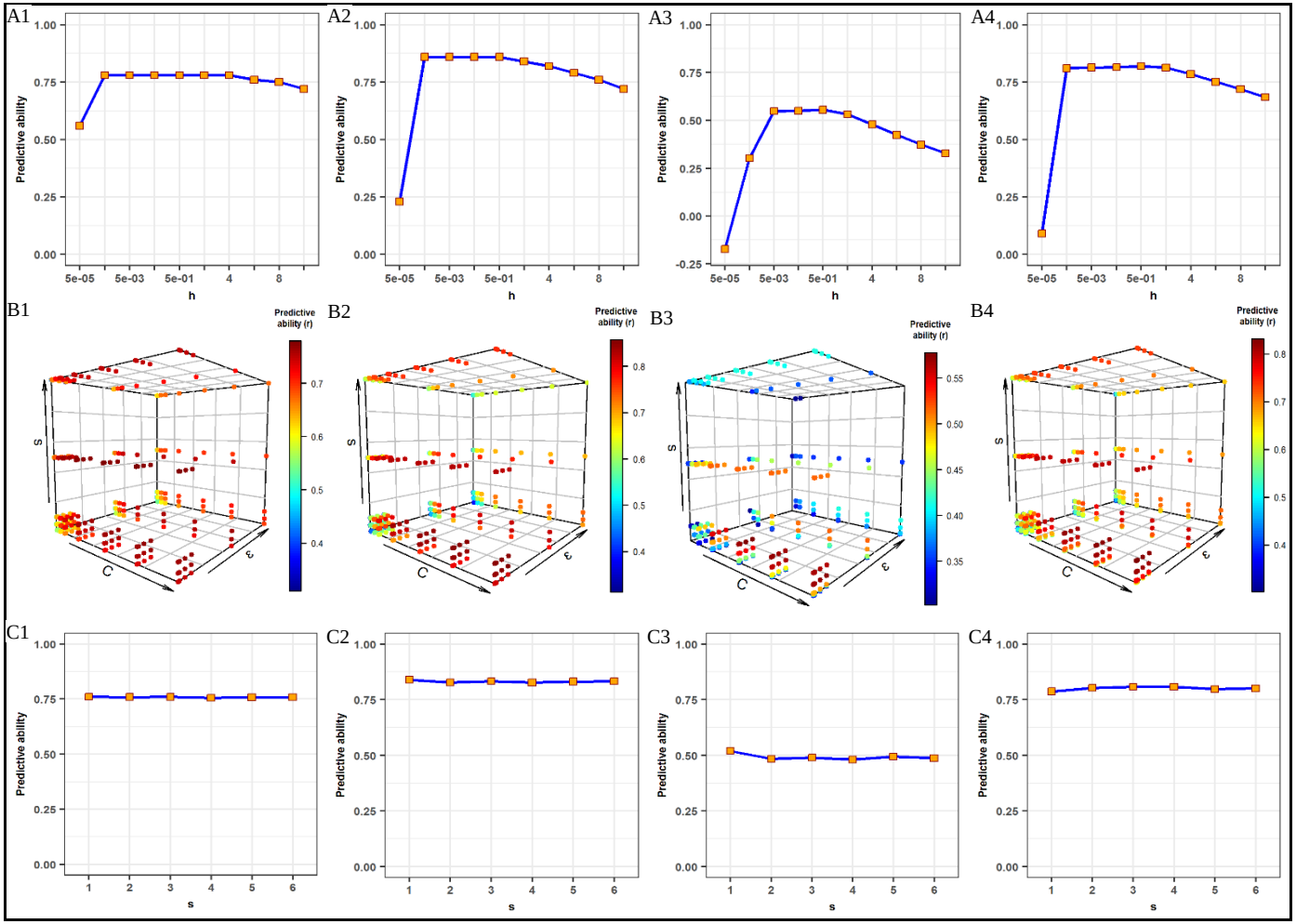


Figure 9. Effect of tuning parameters on the predictive ability of three nonparametric genomic prediction models. A: Effect of bandwidth parameter (h) on predictive ability of RKHS; B: Effect of ε-insensitivity, complexity parameter (C) and bandwidth parameter (σ) on predictive ability of SVR (Note: Only combinations providing predictive ability greater than 0.3 are shown to avoid overcrowding); C: Effect of number of neurons (s) on predictive ability of NN. Vertical planes (from left to right) represent 1. GY (dataset I), 2. PH (dataset I), 3. GY (dataset II) and 4. PH (dataset II)


A1
A2
A3
A4
B1
B2
B3
B4
C1
C2
C3
C4
Predictive ability
h
s
C
ε
Predictive ability (r)

Chapter 4: Optimization of Training Set Composition for Genomic Prediction of Early-Stage Single Crosses in Maize

The promising results of genomic prediction of single-cross performance indicate the possibility for evaluation of inbreds based on performance in hybrid combinations from the beginning. The early-stage prediction of single-cross performance has advantage that all possible single crosses between the inbreds can be evaluated and the commercial hybrid development is hastened. The information, however, is lacking about optimal construction of training set (TRS) for genomic prediction of early-stage single crosses. The objectives of the present study were to investigate the effect of different levels of relatedness on the predictive ability, usefulness of deterministic formula to estimate the accuracy in advance and potential for TRS optimization based on prediction error variance (PEV) and coefficient of determination (CD) criteria. We used 481 single crosses between the randomly selected 89 recombinant inbred lines (RILs) belonging to Iowa Stiff Stalk Synthetic (BSSS) and 103 RILs belonging to Non-Stiff Stalk Synthetic (NSSS) heterotic group. The main results obtained in this study were 1. Genetic relationship in addition to the number of tested parents of single crosses considerably influenced their predictive ability, 2. Increasing TRS size by combining differently related single crosses improved the predictive ability, 3. Expected prediction accuracies based on PEV agreed well with empirical prediction accuracies when population structure was accounted and 4. Genomic prediction models constructed on TRS optimized with PEV and CD mean criteria provided increased predictive ability than

stratified and randomly sampled TRS. Overall, the results of this study suggest that genomic prediction of early-stage single crosses with TRS optimization using PEV and CD mean criteria would greatly enhance the efficiency of hybrid breeding.

INTRODUCTION

In a typical hybrid maize breeding program 10s to 100s of inbred progenies or doubled haploid lines (DHLs) are generated from a single biparental cross between elite parental lines. The number of biparental crosses varies between breeding programs, but can easily be on the order of hundreds. Early-stage selections have traditionally been based on a combination of *per se* performance and topcross testing, while evaluation of specific new single crosses is performed during the advanced stages (Hallauer and Miranda 1988; Schrag *et al.* 2009). The advantage of topcross testing is that it's an efficient way to identify candidate lines with poor general combining ability early in the pipeline. There are, however, important limitations in this process. First, two or more years of topcross testing can lengthen the time to hybrid commercialization and parent selection (Longin *et al.* 2007). Secondly, all possible crosses among available lines are not evaluated, which leaves open the possibility that some outstanding single crosses are never tested. Early-stage selection of lines based on single-cross performance would therefore desirable. Despite this, early evaluation of lines based on single-cross performance is not often practical because of the typically huge number of possible parental combinations.

Recently both simulation and field studies have indicated that genomic prediction holds great potential for accurately predicting single-cross performance (Technow *et al.* 2012; Massman *et al.* 2013a; Technow *et al.* 2014; Kadam *et al.* 2016). Genomic prediction is based on genotypic and phenotypic data collected on a subset of individuals referred to as the training set (TRS), which is used to train a statistical model for prediction of genetic values of individuals having only been genotyped (Meuwissen *et al.* 2001). The accuracy of genomic prediction is highly dependent on the size and composition of the TRS (Lorenz *et al.* 2011; Clark *et al.* 2012; Lin *et al.* 2014). While the cost and throughput of genotyping large populations has been facilitated by a sharp decrease in the cost of sequencing-based genotyping (Elshire *et al.* 2011), phenotyping for traits such as grain yield remains relatively expensive and comprises a major bottleneck in a breeding program (Araus and Cairns 2014). It is, therefore, important to design an optimal TRS that maximizes prediction accuracy while minimizing TRS size and hence costs associated with phenotyping.

It would be very important for resource allocation in optimizing TRS to know the achievable prediction accuracy from existing germplasm in the breeding program. In an established hybrid maize breeding program, hybrid performance data on recombinant inbred lines (RILs) or DHLs from multiple related and unrelated populations is likely to be available. In this context, previous studies on genomic prediction of topcross performance evaluated different scenarios including training and prediction within a single biparental population, training across multiple biparental populations and prediction within a single biparental population, and training and prediction across

multiple biparental populations (Albrecht *et al.* 2011; Schulz-Streeck *et al.* 2012; Zhao *et al.* 2012; Jacobson *et al.* 2014; Albrecht *et al.* 2014). These studies, however, did not consider the genetic relationships among populations combined into TRS and test set (TS). Knowledge on how prediction accuracy varies when multiple differentially related populations are combined into a single TRS would be useful for optimal TRS construction.

Another approach to optimizing the TRS is to use "selective phenotyping" strategies for designing a TRS. The goal of this approach is to select a minimum number of the most informative individuals for model training. Training set design strategies include functions that minimize the prediction error variance (PEV) or maximize the coefficient of determination (CD) of the TRS. Rincent *et al.* (2012) reported that training sets optimized on PEV and CD mean give higher model prediction accuracy compared to random sampling. Another approach to TRS design includes maximizing the genetic diversity of the TRS by stratified sampling across known genetic groups, or use of clustering algorithms such as *k*-means. Isidro *et al.* (2015) found that CD mean provided highest accuracies in a wheat dataset with mild population structure, but stratified sampling across groups provided the highest accuracies in a rice dataset with strong population structure.

Ability to predict achievable prediction accuracy of a genomic prediction model would help facilitate optimization of models. The question of achievable accuracy has been investigated from a theoretical perspective. Daetwyler *et al.* (2008, 2010) proposed

a formula to determine the expected prediction accuracy as $\hat{r}_{g\hat{g}} = \sqrt{N_T h^2/(N_T h^2 + M_e)}$ where $\hat{r}_{g\hat{g}}$ is the correlation between true and predicted genotypic value, N_T is the TRS size, h^2 is the trait heritability and M_e is the effective number of chromosome segments. The correlations, however, between the expected and empirical prediction accuracies were inconsistent (Combs and Bernardo 2013). In a study by Riedelsheimer *et al.* (2013), the expected and empirical prediction accuracies agreed well in five maize biparental populations, but the correlation between empirical and expected prediction accuracies for any given trait in a single population have also been reported to be very low (Combs and Bernardo 2013; Lian *et al.* 2014). One reason for the inconsistency in agreement between expected and observed prediction accuracies is an inability to accurately estimate M_e (Lian *et al.* 2014). Alternatively, a calculation of the expected accuracy which does not require an estimation of M_e is based on an estimate of PEV output from the mixed models equations (Clark *et al.* 2012) as $\hat{r}_{g\hat{g}} = \sqrt{1 - (PEV/G_{ii}\sigma_a^2)}$ where G_{ii} is the diagonal element of the genomic relationship matrix for individual *i* and σ_a^2 is the additive genetic variance. The usefulness of the above formula, however, has not been tested in the context of genomic hybrid prediction. A unique aspect to the design of a TRS for hybrid prediction is that candidates for inclusion in the TRS are not limited to individuals that had already been created, but rather ecompasses all possible inter-heterotic group crosses between all individuals at any stage in the breeding pipeline. The possibility to infer all possible hybrid genotypes based on inbred progeny genotypes dramatically increases the size of the genetic space from which TRS's can be designed.

The overall aim of this study was to optimize the application of genomic prediction of single-cross performance with a particular consideration to early-stage single crosses. The three specific objectives were to 1) Determine the effect of different levels of relatedness between TRS and TS on the predictive ability of early-stage single crosses; 2) Test the usefulness of deterministic formula based on PEV to predict the achievable accuracy in advance; 3) Compare different TRS optimization strategies for genomic prediction of early-stage single crosses.

MATERIALS AND METHODS:

Germplasm

Inbred progenies representing the early stages of a maize hybrid breeding pipeline for this study were 89 RILs randomly derived from six biparental families belonging to the Iowa Stiff Stalk Synthetic (BSSS) heterotic group, and 103 RILs randomly derived from six biparental families belonging to the Non-Stiff Stalk Synthetic (NSSS) heterotic group. An incomplete factorial was used to create 481 inter-heterotic group single crosses. Parents of the biparental families from which the RILs were derived were Plant Variety Protection expired (ex-PVP) inbred lines. Eight ex-PVP inbreds were used to generate the six biparental families in BSSS heterotic group and six ex-PVP inbreds were used to generate the six biparental families in the NSSS heterotic group. The parents of biparental families, number of RILs in each biparental family and number of single crosses per family-wise cross combinations are listed in Table 14. The RILs were derived from F5 selfing generation. The mean of single crosses per BSSS and NSSS RIL was 5.4

(ranged 1-14) and 4.7 (ranged 1-8). Field trials were carried out in Nebraska during 2014 (Locations: Mead and York) and 2015 (Locations: Havelock and York). The trials included 450 single crosses in 2014 and 467 single crosses in 2015, with 436 single crosses included in both years. The total number of single crosses evaluated in at least one year was 481. The four location-year combinations were considered as separate environments. The experimental design was randomized incomplete block with two replications at each environment. The single cross entries were grown in two row plots of dimensions 4.46 m in length and 0.76 m in width. The target planting density was 88506 seeds per hectare. Phenotypic data were recorded on several agronomic traits of which plant height (PH) and grain yield (GY) were used for this study. PH was measured from base of the plant to collar of flag leaf at post anthesis stage. Six plants (three per row) were randomly chosen for PH measurement. The average of six measurements formed PH data for of that particular plot. GY was converted to $MTha^{-1}$ on a 155 g kg^{-1} moisture basis. The GY data on plots having more than 10 plants lodged were discarded.

Genotyping by Sequencing

Deoxyribonucleic acid (DNA) was extracted from lyophilized leaves using the Qiagen DNeasy Plant 96 kit. Briefly, five seeds of each RIL/DHL were planted in greenhouse for leaf sample collection. A pooled leaf tissue was sampled from the five seedling and immediately frozen in liquid nitrogen. The leaf samples were lyophilized and later used for DNA extraction. The library preparation and sequencing were performed at Institute for Genomic Diversity (IGD) at Cornell University as described by

Elshire *et al.* (2011). Single nucleotide polymorphisms (SNPs) were called from the raw sequence data using the TASSEL GBS Pipeline version 3.0 (Glaubitz *et al.* 2014). SNPs were filtered for maximum missing percentage (> 20%) and minimum minor allele frequency (< 5%) wherein heterozygotes were treated as missing data. Missing data was replaced by the mean value for the markers (i.e. naïve imputation). Of the markers remaining after filtration, markers that were polymorphic among both BSSS and NSSS lines were retained for further analysis. A total of 23923 SNPs were thus obtained. The marker profiles of single crosses were deduced from their parental SNP information.

Phenotypic Data Analysis

The analysis of phenotypic data across the environments was performed using the following statistical model

$$y_{iklq} = \mu + g_i + e_k + (ge)_{ik} + r_{l(k)} + b_{q(kl)} + \varepsilon_{iklq} \quad \text{... (1)}$$

where y_{iklq} is the phenotypic observation for i^{th} single cross evaluated in the k^{th} environment in the l^{th} complete block (i.e. replicate) and q^{th} incomplete block. Other terms in the model were as follow: μ is the grand mean; g_i is the effect of i^{th} single cross; e_k represents the effect of the k^{th} environment; $(ge)_{ik}$ represents the interaction effects of i^{th} single cross with k^{th} environment; $r_{l(k)}$ represents the effect of the l^{th} complete block nested within the k^{th} environment; $b_{q(kl)}$ represents the effect of the q^{th} incomplete block nested within the l^{th} complete block in the k^{th} environment; and ε_{iklq} represents the residual. All the effects were considered as random except environment and replication

nested within environment effects were modeled as fixed effects. The distributions of g_i and $(ge)_{ik}$ were assumed as follow: $g_i \sim N(0, I\sigma_g^2)$ and $(ge)_{ik} \sim N(0, I\sigma_{g\times e}^2)$ respectively. Error and incomplete block variances were specified to be heterogeneous among the environments. Stand count was included as a covariate in model (1) for analysis of GY which reduced the error variance by four percent.

All the variance components were estimated using restricted maximum likelihood estimates (REML) procedure as implemented in ASReml-R software (Butler *et al.* 2009). Significance of the variance components was tested using likelihood ratio tests at 0.001 level of significance. The entry-mean heritability of each trait was estimated as per the following formula: $H^2 = \sigma_g^2 / \left(\sigma_g^2 + \frac{\sigma_{g\times e}^2}{e} + \frac{\sigma_e^2}{re}\right)$ where, e is the number of environments, and r is the number of replications in each environment.

Genomic Prediction Model

We used genomic best linear unbiased prediction (GBLUP) model for prediction of single-cross performance as:

$$y = 1_n\mu + Z_f g_f + Z_m g_m + Zs + e \quad \text{..... (2)}$$

where, y vector of single-cross BLUPs obtained from equation (1); g_f is GCA effect of female (BSSS RIL), g_m is the GCA effect of male (NSSS RIL) and s is the SCA effect of cross; Z_f, Z_m and s were the corresponding design matrices. All the effects were modelled as random with covariance structure as follow: $g_f \sim N(0, G_f\sigma_f^2)$, $g_m \sim N(0, G_m\sigma_m^2)$ and $s \sim N(0, S\sigma_s^2)$ where σ_f^2, σ_m^2 and σ_s^2 were the variances of GCA

effects of females, males and SCA effects, G_f, G_m and S were the additive genomic relationship matrices among females, males and dominance genomic relationship matrix among single crosses correspondingly. G_f, G_m were calculated according to VanRaden (2008) as $G_f(or\ G_m) = \frac{WW'}{2\sum_{k=1}^{p} p_k(1-p_k)}$ where W is an $n \times m$ matrix with $w_{ij} = x_{ij} - 2p_j$ and x_{ij} is marker genotype of i^{th} individual for j^{th} marker which was coded as 0, 1 or 2 for homozygous for major allele, heterozygous and homozygous for minor allele states respectively. The positive elements of $G_f(or\ G_m)$ indicate that corresponding pair of females (or males) have above average relationships, while negative elements indicate that corresponding pair of females (or males) have below average relationships. S was calculated as per Bernardo (2002). The GBLUP model was implemented using ASReml-R software to obtain BLUPs of GCA and SCA effects (Butler *et al.* 2009).

Cross-Validation Schemes

To evaluate the effects of different levels of relatedness between TRS and TS on the predictive ability for single crosses, the RILs were classified into three groups within each heterotic group depending upon their realized genomic relationships (Figure 10). The three BSSS RIL (female) groups were denoted as GF_1, GF_2 and GF_3 and three NSSS RIL (male) groups were denoted as GM_1, GM_2 and GM_3. RILs within a group were more closely related than RILs between the groups. Also, the groups within each heterotic group were differentially related with one group having relatively closer relationship with one of the two other groups and distant relationship with the remaining group. Similarly,

the 481 single crosses between BSSS and NSSS RILs were classified into nine groups with each group containing single crosses between one particular group of BSSS and NSSS RILs. The groups of single crosses, therefore, had different levels of relatedness. Cross-validations were performed so as to predict single crosses within one group at a time (i.e., leave-one-group cross-validation) (Figure 11A). The remaining eight single-cross groups were distinguished into closely and distantly related to validation group depending upon the level of genetic relatedness between the parental RILs of validation group and parental RILs of the remaining groups of single crosses. The letters W, C and D were used to denote the validation group (closest relationship), groups closely related to validation group and groups distantly related to validation group, respectively. It is important to note that relationships are defined here in relative terms and distantly related RILs may be fairly related with each other. The TRS's were constructed with single crosses belonging to closely related (C), distantly related (D) or both (C+D) groups. To evaluate the importance of having single crosses from a validation group, TRS were also formed by combining the single crosses from validation group with single crosses belonging to closely (i.e. C+W), distantly (i.e. D+W) related or both groups (i.e. C+D+W). The effects of above mentioned TRS compositions (i.e. D, C, C+D, C+W, D+W and C+D+W) on the predictive ability were investigated in a scenario when both female and male parental RILs of validation group were tested (G2) and when only female or male parental RILs of the validation group were tested (G1). The predictive abilities were estimated for each group as the Pearson's correlation coefficient between observed single-cross performance and predicted single-cross performance. The mean

predictive abilities over nine single-cross groups for each of the twelve combinations (i.e. six TRS compositions (i.e. D, C, C+D, C+W, D+W and C+D+W) under two scenarios (i.e. G2 or G1)) were used to investigate the effect of genetic relationship on the predictive ability.

Expected Prediction Accuracy Using PEV-Based Deterministic Formula

The expected prediction accuracy was calculated using the formula $\hat{r}_{g\hat{g}} = \sqrt{1 - (PEV/G_{ii}\sigma_a^2)}$ where $\hat{r}_{g\hat{g}}$ is the correlation between true and estimated genetic values, G_{ii} is the diagonal element of genomic relationship matrix for the i^{th} single cross and σ_a^2 is the additive genetic variance. Three different levels of relatedness between TRS and TS (i.e. TRS compositions D, C+D and C), as well as when TRS and TS were randomly sampled across the entire population of 481 single crosses, were considered. TRS of sizes 75, 100, 125 and the maximum possible were used for different TRS compositions. Sampling of TRS out of maximum possible was replicated 10 times for TRS sizes 75, 100 and 125. Empirical and expected prediction accuracies were calculated without and with accounting for population structure based on the following statistical models:

1a. *Expected prediction accuracy without accounting for population structure*

$$\mathrm{y} = 1_n\mu + \mathrm{Z}h + \mathrm{e} \qquad ...(3)$$

1b. *Expected prediction accuracy accounting for population structure*

$$\mathrm{y} = 1_n\mu + \mathrm{W}\mathrm{p} + \mathrm{Z}h + \mathrm{e} \qquad ...(4)$$

2a. *Empirical prediction accuracy without accounting for population structure*

$$y = 1_n\mu + Z_f g_f + Z_m g_m + Zs + e \quad ...(5)$$

2b. *Empirical prediction accuracy accounting for population structuire*

$$y = 1_n\mu + Wp + Z_f g_f + Z_m g_m + Zs + e \quad ...(6)$$

where h is the random genetic effect of single crosses having co-variance structure $h \sim N(0, G\sigma_a^2)$ with G as the additive genomic relationships matrix among single crosses calculated according to method 1 of VanRanden (2008) using single cross genotypes inferred from parental SNP scores; Z is the incidence matix for the single crosses, p is the fixed effect of single-cross group; W is the incidence matrix for the effect of single-cross groups. All the other terms were as descibed in equation (2). The PEV required to calcualate the expected prediction accuracy without and with accounting for population structure was estimated as square of the standard error of BLUP for single crosses i.e. $PEV = var(\hat{h}_i - h)$ obtained with equation (3) and equation (4) respectively. Similarly, the empirical prediction accuracies without and with accounting for population structure were calcuated based on GCA estimated from equation (5) and equation (6) respectively. The empirical prediction accuracy is defined as the correlation between observed and prediction single-cross performance divided by square root of heritability (Dekkers 2007).

Optimization Criteria and Algorithm

Four methods were evaluated for optimizing the TRS which include PEVmean, CDmean, stratified sampling and random sampling. The PEVmean method optimizes the TRS by minimizing the mean of the PEV of the contrast between each TS individual and mean of the TRS individual. The CDmean method optimizes the TRS by maximizing the mean of the CD of the contrast between each TS individual and mean of the population. The CD is defined as the squared correlation between true and predicted contrast of genetic values (Laloe 1993). Its value always lies within a unit interval, a CD value of 0 means that the prediction of the contrast is unreliable and CD value of 1 means that prediction is most reliable. The CD takes into account PEV as well as genetic variance of individuals in TRS. In case of stratified sampling, TRS individuals were sampled from each of the nine groups of single crosses in proportion to their sizes and in case of random sampling, the choice of TRS individuals was entirely random. The statistical equations to calculate PEV and CD were provided by Rincent *et al.* (2012) and Isidro *et al.* (2015).

The genetic algorithm implemented in the R package STPGA (Akdemir 2017) was used for TRS optimization with PEVmean and CDmean. Briefly, genetic algorithms are stochastic search algorithms which solve optimization problems using evolutionary strategies. At each iteration, the algorithm randomly exchange one genotype between solution set and the remaining set and fitness function is subsequently used to determine if the genotype is accepted. The population of solution thus evolve towards better solution by keeping only elite genotypes at each iteration. The genetic algorithm in

STPGA is specialized for subset selection and has two additional features, tabu (memory) and “inference through prediction based on current population of solutions” which the author named as LA-GA-T (look ahead genetic algorithm with tabu) algorithm. Tabu search keeps track of previously tried solutions to exclude the inferior solutions in further search. Additionally, an ideal estimated solution to search is formed by regressing the fitness of existing solutions on their coding. This algorithm was recently suggested for selecting individuals for TRS in GS which is a combinatorial optimization problem (Akdemir *et al.* 2015). In our implementation, STPGA default settings were used.

To evaluate the four optimization methods, the 481 single crosses were divided into 10 mutually exclusive folds. One fold was used as a TS and remaining nine folds were combined to form TRS candidate set. TRS of sizes 50, 100, 150, 200, 250, 300, 350 and 400 were selected from candidate set based on the four optimization methods. This is repeated so that each fold comprised the TS exactly one time. The entire process was replicated twice, resulting in 20 combinations of TRS and TS for each optimization method. The predictive ability was calculated for each combination of TRS and TS and mean predictive abilities over twenty combinations were used to compare the four optimization methods.

The four methods were subsequently investigated for TRS optimization with all possible 9167 single-cross combinations. For this scenario, TS of 50 single crosses was randomly sampled out of total 9167 possible single cross-combinations. TRS of sizes 100, 200, 300 and 400 were selected from the TRS candidate set using four optimization

methods. This procedure was repeated to obtain the 20 combinations of TRS and TS for each method. As phenotypes were not available for all single-cross combinations, we couldn't calculate the empirical predictive ability in this scenario. Instead, we calculated the PEV for each of twenty TSs when models were built on optimal TRS obtained with four methods. The PEV was calculated based on equation (3) as (Isidro *et al.* 2015): $PEV = (Z'MZ + \lambda G^{-1})^{-1} \times \sigma_e^2$ where $M = I - 1_n(1_n'1_n)^{-}1_n'$, and $\lambda = \frac{\sigma_e^2}{\sigma_g^2}$. To compare the advantage of TRS optimization with all 9167 single cross-combinations against 481 single crosses, TRS of sizes 100, 200, 300 and 400 were also selected from the TRS candidate set comprising of 481 single crosses. The mean PEV was calculated over the twenty TRS obtained with all 9167 single cross-combinations and TRS obtained out of 481 single crosses. Finally, we calculated the expected prediction accuracies using a PEV-based deterministic formula for different TRS and TS combinations.

RESULTS

Variance Components and Heritabilities

Genetic variation between single crosses was significant at $\alpha = 0.01$ in the entire population for both GY and PH, as well as within each of the nine groups of single crosses (Table 15). Single crosses made by crossing RILs from the same biparental families substantially differed for grain yield, with ranges expressed as a percent of the group mean being at least 23% and up to 35%. Also, considerable amount of genetic variation was present between the groups of single crosses. The proportions of genetic

variation between groups to the total genetic variation were 36 and 9 percent for GY and PH respectively. The broad sense heritabilities were moderate to high for GY and high for PH in different single-cross populations.

Single-Cross Predictive Abilities with Different Levels of Relatedness

The effect of different levels of relatedness between TRS and TS on the predictive ability was evaluated by constructing the TRS with single crosses belonging to closely (C), distantly (D) related and both groups (C+D) under a scenario when both female and male parental RILs of TS group of single crosses were tested (G2) and when only female or male parental RILs of TS group of single crosses were tested (G1). The mean (over TRS compositions D, C and C+D) predictive abilities for the G2 scenario were considerably higher than G1 scenario (Figure 11B). The mean increments in the predictive abilities for G2 scenario over G1 scenario were 40 and 33 percent for GY and PH, respectively. Higher predictive abilities were obtained when TRS was composed of single crosses belonging to closely related groups (C) than distantly related groups (D). Specifically, for the G2 scenario, predictive ability was 28 percent higher for GY when a TRS was composed of C + W (closest relationship) compared to a TRS was composed of D (distant relationship). The difference for the same comparison was 13 percent for plant height. Combining closely and distantly related single-cross groups (i.e., C+D) into a single TRS increased predictive abilities. When single crosses from the validation group were added to the TRS, predictive abilities were only slightly increased for G2 scenario (5 and 4 percent for GY and PH, respectively), while a substantial improvement was

observed for the G1 scenario (32 and 31 percent for GY and PH, respectively). Predictive abilities from GCA-only models were very similar to models including both GCA and SCA. Empirical predictive abilities for single crosses henceforth were based on only GCA effects estimated from equation (2).

Comparisons of Expected and Empirical Prediction Accuracies

Expected and empirical prediction accuracies were compared when the TRS was composed of single crosses belonging to subsets D, C+D and C, as well as when the TRS and TS were randomly sampled across the entire population of 481 single crosses. TRS of sizes 75, 100 and 125 and total possible were constructed for these TRS compositions. Expected and empirical prediction accuracies were closely related both without and with accounting for population structure when TRS and TS were randomly sampled (Figure 12). However, when the TRS was composed of D, C+D and C groups, the expected prediction accuracies without accounting for population structure were generally upwardly biased (Figure 12A and 12B). The upward bias was highest for D followed by C+D and low for C. These trends between expected and empirical prediction accuracies were consistent across traits but more pronounced for GY compared to PH. When the population structure is accounted for by modelling single-cross group effect in the model (equation 4 and 6), the correspondence between expected and empirical prediction accuracies for C, D and C+D greatly increased (Figure 12C and 12D).

Comparisons of Training Set Optimization Methods

The PEVmean and CDmean methods of TRS optimization provided comparable and highest predictive abilities over stratified and random sampling method which obtained similar and lowest predictive abilities (Figure 13). The difference in predictive abilities of these two groups of TRS optimization methods was pronounced at lower TRS sizes. This trend was persistent across the traits. Specifically, the mean difference in predictive abilities between these two pairs of methods were 8 and 28 percent at TRS size 50, and 3 and 5 percent at TRS size of 200 for GY and PH, respectively. About one hundred more single crosses were needed when TRS was constructed with stratified and random sampling compared to when TRS was constituted based on PEVmean and CDmean method to obtain the predictive abilities equivalent to maximum (0.76 ± 0.02 for GY and 0.86 ± 0.02 for PH).

Training Set Optimization with All Possible Single Crosses

The four methods of TRS optimization were subsequently applied to optimize TRS with all possible 9167 single crosses between 89 females and 103 males. The PEV was lowered when all possible single-cross genotypes were used for TRS optimization with PEVmean and CDmean method compared to when selecting out of only 481 single crosses (Figure 14). The mean decrease in PEV across different TRS sizes was about 21 percent. The benefit of TRS optimization was higher with all possible single crosses was higher at larger TRS sizes. In case of stratified and random sampling, similar PEV was

obtained when optimizing TRS with all possible single crosses against with only 481 single crosses.

DISCUSSION

Genomic prediction of single crosses in the early stages is attractive because it reduces the time required for hybrid development by skipping topcross testing. Moreover, it allows the evaluation of genetic values of all possible single-cross combinations which could include some potential single crosses that may have dropped because of chance poor performance with elite tester. In the present study, we studied TRS optimization for genomic prediction of single crosses with main consideration to early-stage single crosses. Specifically, we investigated three factors relevant for TRS optimization which included the effect of different levels of relatedness between TRS and TS on the predictive ability, usefulness of a PEV-based deterministic formula to estimate the accuracy in advance and potential for TRS construction based on PEVmean and CDmean criteria.

Effect of Different Levels of Relatedness on the Predictive Ability

The construction of TRS for an individual biparental population is ideal as high predictive abilities can be ensured under this scenario due to a close genetic relationship between TRS and TS (Lorenzana and Bernardo 2009; Zhao *et al.* 2012). It, however, has a major limitation that subset of individuals from each population needs to be phenotyped, increasing the operating cost of genomic selection. Also, individual population sizes needs to be sufficiently large to reliably perform within population

predictions (Schulz-Streeck *et al.* 2012). With these considerations, it would be advantageous if the performance of genotypes within a population could be predicted by using phenotypic data on related populations.

The early-stages of maize hybrid breeding consist of generation of RILs or DHLs from several biparental populations for testing into hybrid combinations. It is, therefore, likely that hybrid performance data from differently related populations is available. The available data could be efficiently used if the effect of different levels of genetic relatedness between TRS and TS on the predictive ability is understood. To investigate this, we varied the composition of TRS to create different levels of relatedness. In total, we investigated the single-cross predictive abilities under twelve levels of genetic relatedness between TRS and TS. The results of this analyses helped to better understand the effect of genetic relationship and TRS size in combination with number of tested parents of single crosses on the predictive ability for single crosses. When no phenotypic data was available for single crosses within a group, considerably higher predictive ability was obtained when both female and male parental RILs of that particular single-cross group were tested in single crosses than when only female or male parental RILs were tested. This results is analogous to the trends in the predictive abilities of T2, T1 single crosses reported from the previous studies (Technow *et al.* 2012, 2014; Massman, *et al.* 2013a; Kadam *et al.* 2016). Nevertheless, we also found that the predictive abilities could be lower even when both the parental RILs were tested. This can be evidenced from the lower predictive ability for single crosses within a group when TRS was constituted from single crosses belonging to only distantly related groups (D) under G2

scenario. This indicates that the genetic relationship needs to be considered in addition to number of tested parents of single crosses to obtain good predictive ability.

Further, the single crosses within a group were predicted with a greater accuracy by combining the single crosses belonging to closely and distantly related groups into TRS than using only the single crosses belonging to closely related groups. Previous studies on genomic prediction of topcross performance have reported an increase (Albrecht *et al.* 2011; Schulz-Streeck *et al.* 2012; Zhao *et al.* 2012) as well as a decrease (Zhao *et al.* 2012) in the predictive ability when the TRS is composed of multiple populations than when TRS composed of individulas from the same the population as TS individuals. The variable effects of combining multiple populations into TRS may be explained on the following two considerations. If present, factors such as marker x population interaction, epistasis and opposite linkage phases between marker and QTL in different populations can negatively impact the predictive ability. In the absence or negligible presence of these factors, however, the predictive ability can be higher compared to within population prediction due to potential increase in the TRS size.

Usefulness of Deterministic Formula to Design Training Set

Methods or formula to calculate the expected prediction accuracy are desired to guide breeders about optimal design of a TRS for achieving a certain level of accuracy for selection. Previous studies in maize indicated the limited usefulness of deterministic formula proposed by Daetwyler *et al.* (2008, 2010) due to ambiguity in the approximation of M_e parameter (Combs and Bernardo 2013; Lian *et al.* 2014). Instead,

the factor $r^2(Nh^2)^{1/2}$, where r^2 is the mean LD between markers, was shown to have moderate and statistically significant association with empirical predictive ability (Jacobson *et al.* 2014; Lian *et al.* 2014).

In the present study, we tested the usefulness of PEV obtained from the mixed model equations to calculate an expected prediction accuracy under different levels of genetic relationships. The advantage of this method is that it doesn't require M_e parameter and accuracy is obtained in conjunction with predictions without the need of separate calculations. We found good correspondence between the expected and empirical prediction accuracies. However, the expected accuracies were upwardly biased when population structure was not accounted and when TRS composed of less related single crosses (Figure 12A and 12B). When the TRS and TS were randomly sampled across the entire population, the expected and empirical prediction accuracies were closely correlated both without and with accounting for population structure. This suggest that upward bias in case of D, C and C+D may be due to different population structure of TRS and TS. To test this, we calculated the expected and empirical prediction accuracies taking into account the population structure (equation 4 and 6). The correspondence between expected and empirical prediction accuracies thus obtained for D, C and C+D as well as for random sampling of TRS and TS was become very high (Figure 12C and 12D). Overall, these results suggest that the accuracy of genomic prediction of single crosses can be reliably predicted in advance using the PEV-based deterministic formula.

Selection of Single-Cross Combinations for Phenotyping

Given the large number of possible single crosses between available inbreds, the choice of single crosses for phenotyping is not straightforward. Previous studies on genomic prediction of single-cross performance highlighted the different criteria for TRS construction. These criteria include number of tested parents of a single crosses (i.e. 2, 1 and 0) and number of single crosses per tested parent. The results indicated that prediction accuracies increases with increase in the number of tested parents (Massman *et al.* 2013a; Technow *et al.* 2014; Kadam *et al.* 2016) and number of single crosses per tested parent (Technow *et al.* 2014). Nevertheless, detailed information on how to select the single crosses for phenotyping is lacking. Importantly, above mentioned criteria do not take into account the genetic relationships among the inbreds which can greatly influence the predictive ability. In that context, the question arises whether both parents of single crosses needs to be tested if the inbreds are closely related vs distantly related. Moreover, the results of this study shows that having both parents tested doesn't necessarily provide high prediction accuracies especially if TRS and TS are distantly related (comparable to G2 scenario with TRS containing single cross belonging to group D). Also, it is unclear how many single crosses per parent should be made to achieve certain level of accuracy. Overall, it appears that different criteria are needed for efficiently designing TRS for genomic prediction of single crosses.

The TRS optimization criteria based only on genotypic data of single crosses would be ideal because genotypic data of all possible single crosses can easily be derived from the marker profiles of parental inbreds. In this context, we evaluated the PEVmean

and CDmean criteria for selecting single crosses for TRS construction. The PEVmean criteria minimizes the mean PEV of contrast between each TS genotype and mean of the TRS. The CDmean criteria maximizes the mean of CD of contrast between each TS genotype and mean of the population. These TRS optimization criteria require only genotypic data. Our results indicate that TRS construction based on PEVmean and CDmean methods provide higher predictive abilities over TRS construction based on stratified and random sampling (Figure 13). The advantage PEVmean and CDmean methods for TRS construction increased when all possible single-cross combination were considered. In spite of the huge number of single-cross combinations to select from, PEVmean and CDmean methods obtained TRS providing precision comparable to when selecting from just 481 single crosses. In contrast, the precision of stratified and randomly sampled TRS were severely decreased (Figure 13). The mean expected prediction accuracies using TRS of size 200 selected with PEVmean and CDmean methods were equivalent to mean expected prediction accuracies using stratified and randomly sampled TRS of size 400 (Table 16). This indicate a great advantage for PEVmean and CDmean methods over stratified and random sampling for optimal TRS design.

Back to Shull's "Pure Line Method of Corn Breeding"

"Pure line method of corn breeding" proposed by Shull (1909) involves development of inbred by self-pollination and subsequently their evaluation in single-cross combinations. Early evaluation of inbreds based on single-cross performance is

difficult because of a large number of single crosses possible among the available inbreds. In a commercial hybrid development, therefore, early selection of inbreds are performed on the basis of per se and topcross performance while selection based on single-cross performance are delayed until advanced stages. This process has disadvantages in that all possible single-cross combinations among available inbreds cannot be evaluated and the time required for commercial hybrid development increases. The promising results of genomic prediction of single-cross performance obtained in recent studies offers a possibility for direct implementation of Shull's idea (Massman *et al.* 2013a; Technow *et al.* 2014; Kadam *et al.* 2016). For the successful implementation of genomic prediction in the early stages, the optimal construction of TRS is critically important. The results of present study suggest a great scope for optimal TRS design using deterministic formula and optimization criteria. If the data is available from multiple differently related material, PEV based estimate of accuracy would enable the hybrid maize breeder to determine the achievable accuracy from existing related material and if there is a need for new phenotyping. The choice of single crosses to make and phenotype should be determined based on PEVmean or CDmean criteria. Early-stage genomic prediction of single crosses using optimally constructed TRS holds potential to increase the efficiency of maize hybrid breeding.

Table 14. Biparental families, number of RILs in each biparental family, number of single crosses for each of thirty six family-wise cross combinations. The total number of single crosses per biparental family are listed in the margins. The inbred groups denotes the classification of RILs based on realized genomic relationship.

Inbred groups	Biparental families	GM1		GM2		GM3		**Total**
		LH109 X LH123 (18)	LH109 X LH59 (20)	LH123 X LH59 (17)	PHG50 X LH123 (20)	PHZ51 X LH123 (14)	PHZ51 X PHG47 (14)	
GF1	LH132 X B73 (16)	17	8	6	26	13	13	**83**
	LH74 X LH132 (14)	11	8	8	8	3	9	**47**
GF2	PHG86 X PHW52 (20)	20	26	15	25	16	19	**121**
	PHJ40 X PHW52 (3)	2	5	2	6	2	0	**17**
GF3	PHB47 X PHW52 (17)	21	23	19	29	18	13	**123**
	PHB47 X PHK29 (19)	15	16	13	19	13	14	**90**
	Total	**86**	**86**	**63**	**113**	**65**	**68**	**481**

Table 15. Mean, range, genetic variance (σ_g^2) and broad sense heritability (H^2) estimates for whole population and nine groups of single crosses for grain yield (GY; Mt/Ha) and plant height (PH; cm)

Single-cross population	n	GY				PH			
		Mean	Range	$\sigma_g^2 \pm SE$	$H^2 \pm SE$	Mean	Range	$\sigma_g^2 \pm SE$	$H^2 \pm SE$
Whole	481	10.97	8.95-13.24	0.71 ± 0.06	0.76 ± 0.02	243.4	203.3-277.9	158.5 ± 10.8	0.95 ± 0.004
G_{SC1}	44	11.39	9.85-12.42	0.68 ± 0.23	0.73 ± 0.08	247.4	227.0-273.0	208.7 ± 47.2	0.97 ± 0.01
G_{SC2}	48	10.26	8.95-11.79	0.33 ± 0.13	0.58 ± 0.11	239.1	212.2-265.3	106.1 ± 24.1	0.93 ± 0.02
G_{SC3}	38	10.95	9.33-12.84	0.26 ± 0.13	0.55 ± 0.14	244.1	218.1-268.5	319.7 ± 77.2	0.96 ± 0.01
G_{SC4}	53	11.16	9.97-13.08	0.64 ± 0.18	0.72 ± 0.06	244.4	229.3-273.4	112.7 ± 23.4	0.95 ± 0.01
G_{SC5}	48	11.00	9.55-12.13	0.37 ± 0.14	0.58 ± 0.10	241.8	203.3-276.7	115.2 ± 25.3	0.94 ± 0.01
G_{SC6}	37	10.36	8.99-11.51	0.39 ± 0.14	0.68 ± 0.09	237.2	212.2-269.6	89.5 ± 22.7	0.93 ± 0.02
G_{SC7}	75	11.48	9.59-12.92	0.18 ± 0.07	0.46 ± 0.11	245.7	217.5-268.5	127.2 ± 22.0	0.95 ± 0.01
G_{SC8}	80	11.85	9.10-13.24	0.29 ± 0.08	0.59 ± 0.08	250.6	225.1-277.9	109.8 ± 18.7	0.94 ± 0.01
G_{SC9}	58	10.95	9.86-12.49	0.28 ± 0.10	0.60 ± 0.10	244.7	214.0-266.4	166.3 ± 33.4	0.95 ± 0.01

Table 16. Mean expected prediction accuracies for grain yield with four training set (TRS) optimization methods when only 481 single crosses used and when all possible 9167 single crosses were considered.

TRS size	481 single crosses				All possible 9167 single crosses			
	PEVmean	CDmean	Stratified	Random	PEVmean	CDmean	Stratified	Random
100	0.72	0.72	0.65	0.65	0.75	0.74	0.66	0.66
200	0.77	0.77	0.73	0.73	0.82	0.82	0.74	0.74
300	0.80	0.80	0.77	0.78	0.85	0.85	0.78	0.79
400	0.81	0.81	0.79	0.80	0.87	0.86	0.81	0.81

Figure 10. Realized genomic relationship among 10A. BSSS inbreds (females) and 10B. NSSS inbreds (males). GF_1, GF_2 and GF_3 denote the three groups of inbreds based on realized genomic relationship within BSSS and GM_1, GM_2 and GM_3 denote the three groups of inbreds based on realized genomic relationship within NSSS

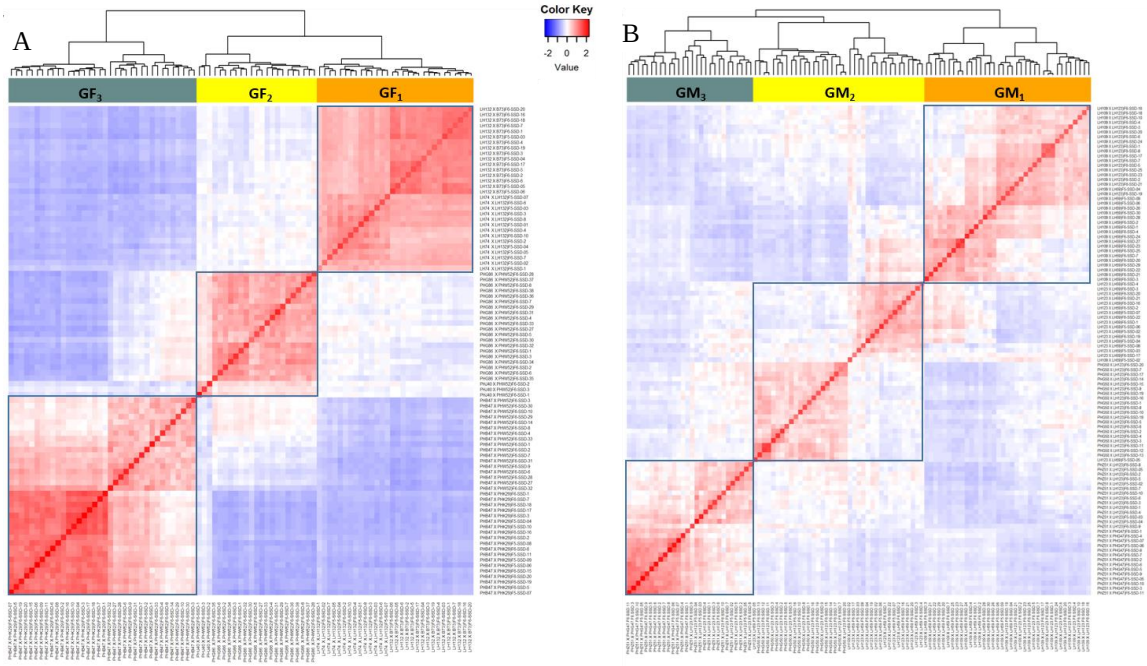


Figure 11. Single-cross predictive abilities under different levels of relatedness between training set (TRS) and test set (TS). 11A.Leave-one-group cross-validation scheme used. GF1, GF2 and GF3 denotes three BSSS inbreds groups and GM1, GM2 and GM3 denotes three NSSS inbred groups based on realized genomic relationships. G2-Both parents of validation group of single crosses (SC) were tested, G1-Either male or female parents of validation group of SC were tested. Letters W, C, and D denotes SC from validation group, closely related group and distantly related group respectively. 11B. Predictive abilities for grain yield (GY) and PH for different TRS compositions using GCA or GCA and SCA.

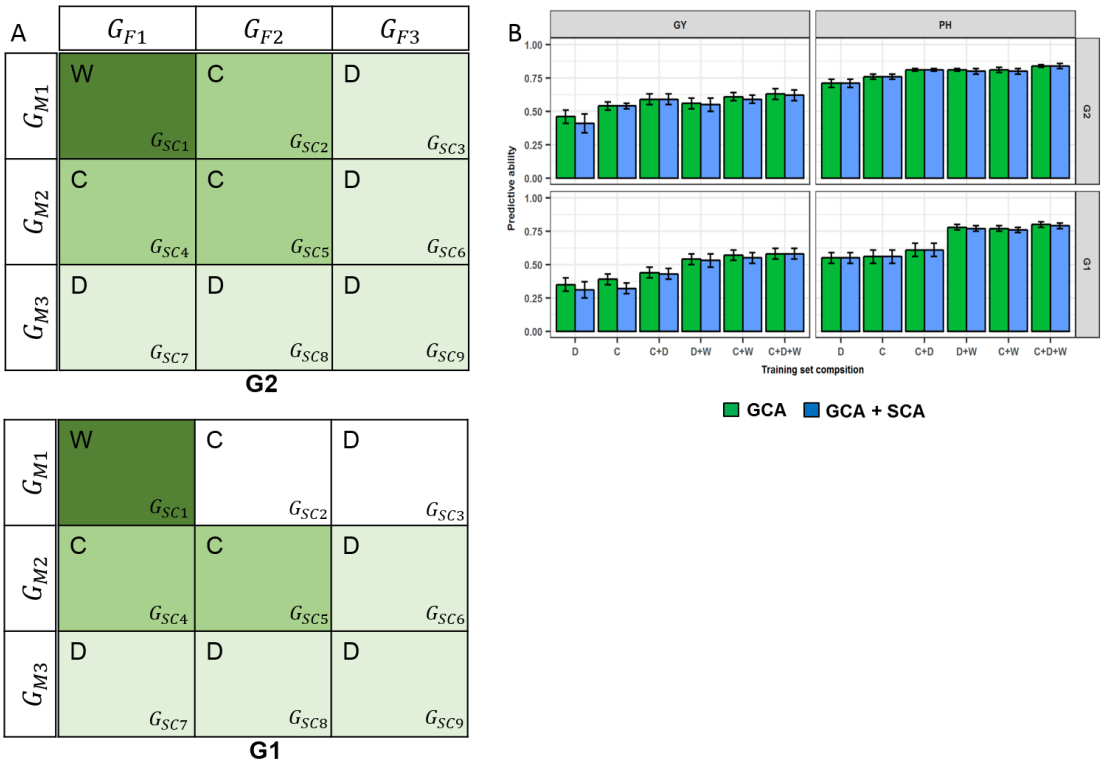


Figure 12. Expected Vs Empirical predictive ability for training set compositions D, C+D, C and across population (R). Without accounting for population structure 12A. Grain yield, 12B. Plant height. With accounting for population structure 12C. Grain yield, 12D. Plant height

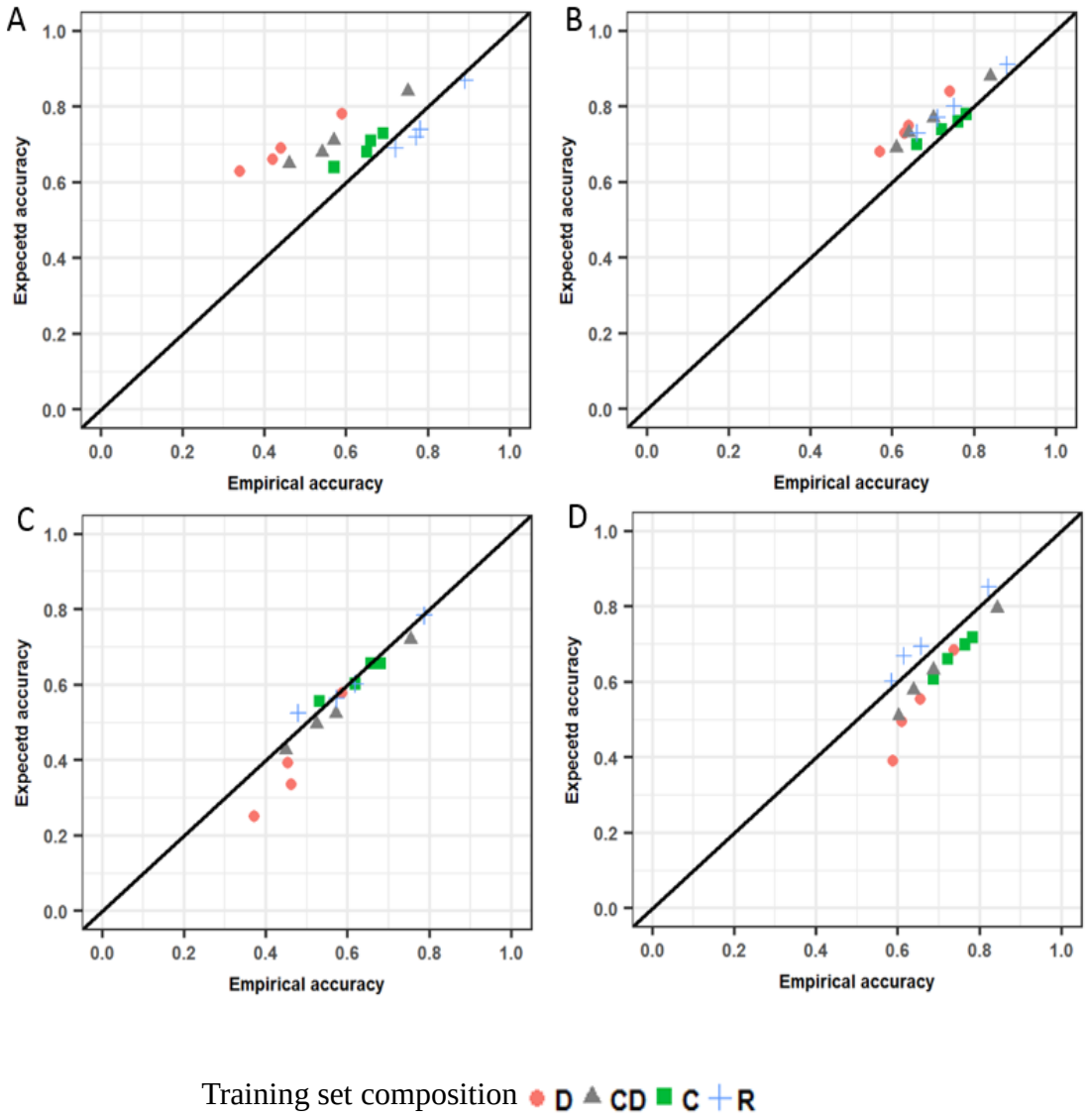


Figure 13. Predictive abilities of model constructed using optimized training set selected out of 481 single crosses based on PEVmean, CDmean, stratified sampling and random sampling. 13A. Grain yield (GY), 13B. Plant height (PH)

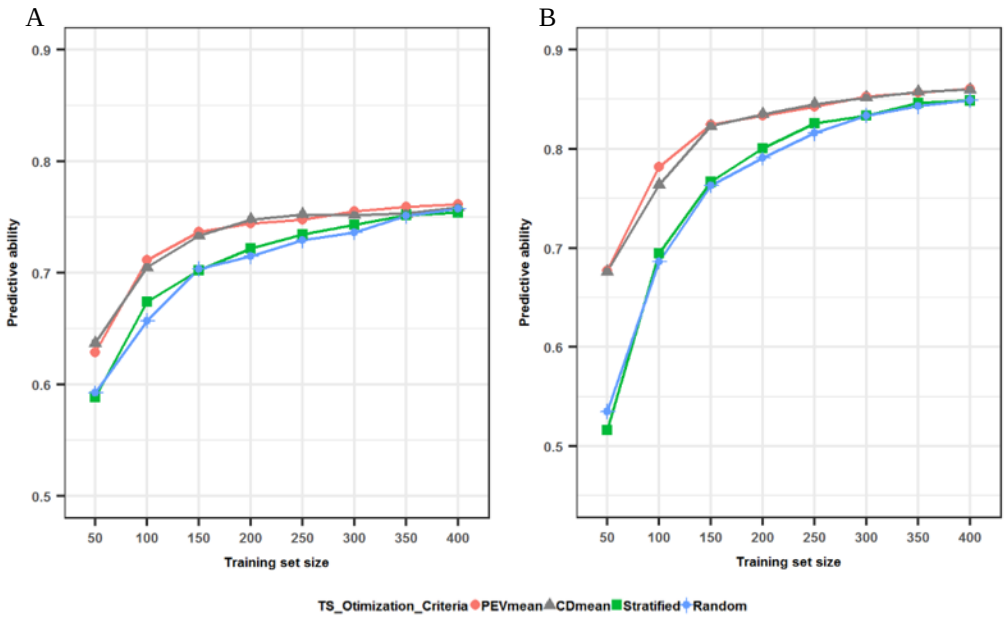
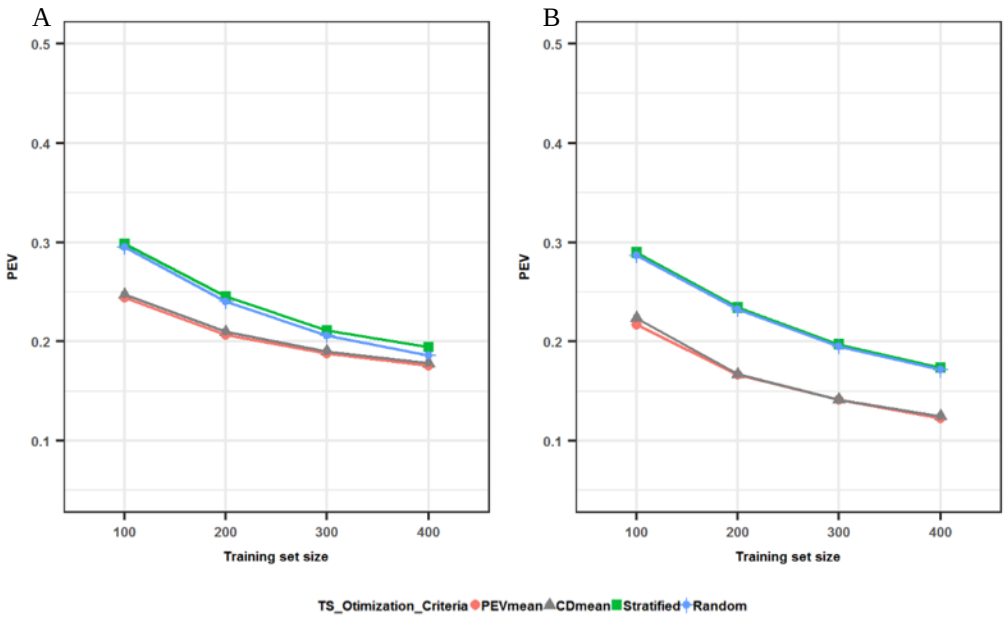


Figure 14. Prediction error variance (PEV) of model constructed using optimized training set selected based on PEVmean, CDmean, stratified sampling and random sampling. 14A. TRS optimization out of 481 single crosses 14B. TRS optimization out of all possible 9167 single crosses



Summary

"Pure line method of corn breeding" proposed by Shull (1909) which involves the development of inbeds by self-fertilization and subsequently their evaluation in hybrid combination is the basic procedure used in hybrid maize breeding. In a typical hybrid maize breeding program, recombinant inbred lines (RILs) or doubled haploid lines (DHLs) are generated from several biparental populations for testing into hybrid combinations. The direct implementation Shull's method at early stages is challenging because of huge number possible single crosses at this stage. To overcome this problem, the early-stage selections of lines for hybrid performance are typically performed on the basis of topcross test using one or multiple testers. Only the lines advanced after topcross stage are evaluated in hybrid combinations. This process has advantage that lines which do not possess good potential for hybrid performance are discarded early in the cycle and resources are concentrated on more promising lines. There are, however, important limitations associated this process. First, additional generations of topcross testing increases the time required for commercial hybrid development, and second, all possible single-cross combinations among available lines cannot be evaluated leaving open the possibility for losing some unique potential single crosses. Therefore, the early-stage selections of lines based on single-cross performance is needed.

In this dissertation, we first investigated the prospects of using genomic selection (GS) approach to predict the single-cross performance at early-stages of hybrid maize breeding using the GS models and training set (TRS) optimization criteria that were not

investigated in previous genomic hybrid prediction studies. The specific objectives of this dissertation were to (1) examine the potential of genomic prediction of single crosses in the early stages of hybrid development pipeline, (2) evaluate the nonparametric genomic selection models for prediction of early-stage single crosses, and (3) optimize the training set composition for genomic prediction for early-stage single-crosses. Two separate datasets were used. Dataset I consisted of 481 single crosses between random set of 89 RILs derived from six biparental families belonging to Iowa Stiff Stalk Synthetic (BSSS) heterotic group and 103 RILs derived from six biparental families belonging to Non-Stiff Stalk Synthetic (NSSS) heterotic group. With similar population structure, dataset II consisted of 312 single crosses between a random set of 46 RILs/DHLs derived from three biparental families belonging to BSSS heterotic group and three biparental families belonging to NSSS heterotic group. Single nucleotide polymorphism (SNPs) data on parental RILs/DHLs in both datasets were obtained using genotyping by sequencing.

The results obtained in this research suggest that genomic prediction holds great promise to identify superior single crosses in the early stages of maize hybrid breeding. The accuracies of genome-based prediction were substantially higher than topcross-based prediction commonly used during the early stages hybrid development. Moreover, genome-based prediction outperformed phenotype-based prediction when only one or none of the parents of single cross were tested. The mean genome-based predictive abilities across GS models and datasets for T2, T1F, T1M, and T0 single crosses were 0.67, 0.60, 0.55, 0.46 for GY and 0.84, 0.74, 0.74, 0.63 for PH correspondingly. Three forms of genomic best linear unbiased prediction (GBLUP) model, one predicted single-

cross performance based on covariance among single crosses estimated from covariance between their parents, other predicted combining abilities of the parents and third predicted genetic effects of single crosses, provided similar accuracies. The results, however, indicated that the relative accuracies three forms of GBLUP could vary depending on the method used for calculation of genomic relationship matrix (GRM) and/or trait predicted.

Interestingly, three nonparametric models namely reproducing kernel Hilbert space (RKHS), support vector regression (SVR) and neural network (NN) did not outperform GBLUP. This results is in contrast to better performance of nonparametric models over GBLUP and ridge regression BLUP (RRBLUP) models in self-pollinating crops such as wheat. The low proportion of specific combining ability (SCA) variance among single crosses in maize and highly heterogeneous nature of maize inbred germplasm could be the reasons for inability of nonparametric models to capture nonadditive effects for single-cross prediction. Consequently, including SCA effect of crosses estimated from GBLUP model and nonparametric models resulted in similar accuracy as predicting based on only general combining ability (GCA) effects.

Finally, the genetic relationship among single crosses in addition to the number of tested parents of single crosses influenced the ability of single-cross prediction. Also, potential beneficial effect of increasing TRS size by combining differently related single-cross groups was observed. Expected prediction accuracy calculated from the prediction error variance (PEV) of mixed model equations agreed well with empirical prediction

accuracy indicating a great scope for TRS design using deterministic formula. Furthermore, the genomic prediction models build on TRS optimized based on PEV mean and coefficient of determination (CD) mean criteria provided higher prediction accuracy than stratified and randomly sampled TRS.

In conclusion, genomic prediction of single crosses in the early-stages of a maize hybrid breeding pipeline holds great potential to redesign hybrid breeding and increase its efficiency. Different models of GS provide comparable prediction accuracy but GBLUP model consisting of combining ability effects of parents may be most appealing for routine use due its simplicity, robustness, maximum accuracy and great importance of GCA and SCA concepts in hybrid breeding. TRS optimization for genomic prediction of single-cross performance at early stages has great scope. Deterministic formula and optimization criteria based on PEV and CD mean should preferably be considered for optimal TRS design.

Bibliography

Akdemir, D., 2017 STPGA: Selection of training populations with a genetic algorithm. bioRxiv 111989.

Akdemir, D., J. I. Sanchez, and J.-L. Jannink, 2015 Optimization of genomic selection training populations with a genetic algorithm. Genet. Sel. Evol. 47: 38.

Albrecht, T., H.-J. Auinger, V. Wimmer, J. O. Ogutu, C. Knaak *et al.*, 2014 Genome-based prediction of maize hybrid performance across genetic groups, testers, locations, and years. Theor. Appl. Genet. 127: 1375–1386.

Albrecht, T., V. Wimmer, H.-J. Auinger, M. Erbe, C. Knaak *et al.*, 2011 Genome-based prediction of testcross values in maize. Theor. Appl. Genet. 123: 339–350.

Araus, J. L., and J. E. Cairns, 2014 Field high-throughput phenotyping: the new crop breeding frontier. Trends Plant Sci. 19: 52–61.

Bauman, L. F., 1982 Review of methods used by breeders to develop superior corn inbreds., in *Proceedings of the... annual corn and sorghum industry research conference American Seed Trade Association, Corn and Sorghum Division, Corn and Sorghum Research Conference*,.

Bernardo, R., 1998 A model for marker-assisted selection among single crosses with multiple genetic markers. Theor. Appl. Genet. 97: 473–478.

Bernardo, R., 1996a Best linear unbiased prediction of maize single-cross performance. Crop Sci. 36: 50–56.

Bernardo, R., 1996b Best linear unbiased prediction of maize single-cross performance given erroneous inbred relationships. Crop Sci. 36: 862–866.

Bernardo, R., 1996c Best linear unbiased prediction of the performance of crosses between untested maize inbreds. Crop Sci. 36: 872–876.

Bernardo, R., 1999a Best linear unbiased predictor analysis. Genet. Exploit. Heterosis Crops 269–276.

Bernardo, R., 2002 *Breeding for quantitative traits in plants*. Stemma Press.

Bernardo, R., 1991 Correlation between testcross performance of lines at early and late selfing generations. Theor. Appl. Genet. 82: 17–21.

Bernardo, R., 1995 Genetic models for predicting maize single-cross performance in unbalanced yield trial data. Crop Sci. 35: 141–147.

Bernardo, R., 1999b Marker-assisted best linear unbiased prediction of single-cross performance.

Bernardo, R., 2003 Parental selection, number of breeding populations, and size of each population in inbred development. Theor. Appl. Genet. 107: 1252–1256.

Bernardo, R., 1994 Prediction of maize single-cross performance using RFLPs and information from related hybrids. Crop Sci. 34: 20–25.

Bernardo, R., 1992 Relationship between single-cross performance and molecular marker heterozygosity. Theor. Appl. Genet. 83: 628–634.

Bernardo, R., 2009 Should maize doubled haploids be induced among F1 or F2 plants? Theor. Appl. Genet. 119: 255–262.

Bernardo, R., and J. Yu, 2007 Prospects for genomewide selection for quantitative traits in maize. Crop Sci. 47: 1082–1090.

Betran, F. J., D. Beck, M. Bänziger, and G. O. Edmeades, 2003 Genetic analysis of inbred and hybrid grain yield under stress and nonstress environments in tropical maize. Crop Sci. 43: 807–817.

Beyene, Y., K. Semagn, S. Mugo, A. Tarekegne, R. Babu *et al.*, 2015 Genetic gains in grain yield through genomic selection in eight bi-parental maize populations under drought stress. Crop Sci. 55: 154–163.

Butler, D. G., B. R. Cullis, A. R. Gilmour, and B. J. Gogel, 2009 Mixed models for S language environments: ASReml-R reference manual (version 3). Qld. Dep. Prim. Ind. Fish.

Calus, M. P. L., A. P. W. De Roos, R. F. Veerkamp, and others, 2008 Accuracy of genomic selection using different methods to define haplotypes. Genetics 178: 553–561.

de los Campos, G., J. M. Hickey, R. Pong-Wong, H. D. Daetwyler, and M. P. L. Calus, 2013 Whole-Genome Regression and Prediction Methods Applied to Plant and Animal Breeding. Genetics 193: 327–345.

Canty, A. J., 2014 EEF. profile 47. Package "boot" 47.

Canty, A., and B. Ripley, 2012 boot: Bootstrap R (S-Plus) functions. R Package Version 1:.

Chang, C.-C., and C.-J. Lin, 2011 LIBSVM: a library for support vector machines. ACM Trans. Intell. Syst. Technol. TIST 2: 27.

Charcosset, A., M. Lefort-Buson, and A. Gallais, 1991 Relationship between heterosis and heterozygosity at marker loci: a theoretical computation. Theor. Appl. Genet. 81:.

Clark, S. A., J. M. Hickey, H. D. Daetwyler, and J. H. van der Werf, 2012 The importance of information on relatives for the prediction of genomic breeding values and the implications for the makeup of reference data sets in livestock breeding schemes. Genet. Sel. Evol. 44: 4.

Cockerham, C. C., 1967 Prediction of double crosses from single crosses. Zücht. 37: 160–169.

Combs, E., and R. Bernardo, 2013 Accuracy of genomewide selection for different traits with constant population size, heritability, and number of markers. Plant Genome 6:.

Comstock, R. E., and H. F. Robinson, 1948 The components of genetic variance in populations of biparental progenies and their use in estimating the average degree of dominance. Biometrics 4: 254–266.

Comstock, R. E., H. F. Robinson, and P. H. Harvey, 1949 Breeding procedure designed to make maximum use of both general and specific combining ability. Agron. J.

Crossa, J., Y. Beyene, S. Kassa, P. Pérez, J. M. Hickey *et al.*, 2013 Genomic prediction in maize breeding populations with genotyping-by-sequencing. G3 Genes Genomes Genet. 3: 1903–1926.

Da, Y., C. Wang, S. Wang, and G. Hu, 2014 Mixed model methods for genomic prediction and variance component estimation of additive and dominance effects using SNP markers. PLoS One 9: e87666.

Daetwyler, H. D., R. Pong-Wong, B. Villanueva, and J. A. Woolliams, 2010 The impact of genetic architecture on genome-wide evaluation methods. Genetics 185: 1021–1031.

Daetwyler, H. D., B. Villanueva, and J. A. Woolliams, 2008 Accuracy of predicting the genetic risk of disease using a genome-wide approach. PloS One 3: e3395.

Davis, R. L., 1927 Report of the Plant Breeder. Rep P. R. Agric Exp Stn 14–15.

De Los Campos, G., D. Gianola, G. J. Rosa, K. A. Weigel, and J. Crossa, 2010 Semi-parametric genomic-enabled prediction of genetic values using reproducing kernel Hilbert spaces methods. Genet. Res. 92: 295–308.

De Los Campos, G., H. Naya, D. Gianola, J. Crossa, A. Legarra *et al.*, 2009 Predicting quantitative traits with regression models for dense molecular markers and pedigree. Genetics 182: 375–385.

Dekkers, J. C. M., 2007 Marker-assisted selection for commercial crossbred performance. J. Anim. Sci. 85: 2104.

Demuth, H., and M. Beale, 2009 Matlab neural network toolbox user's guide version 6. The MathWorks Inc.

Denis, M., and J.-M. Bouvet, 2013 Efficiency of genomic selection with models including dominance effect in the context of Eucalyptus breeding. Tree Genet. Genomes 1–15.

Duvick, D. N., 2001 Biotechnology in the 1930s: the development of hybrid maize. Nat. Rev. Genet. 2: 69–74.

Elshire, R. J., J. C. Glaubitz, Q. Sun, J. A. Poland, K. Kawamoto *et al.*, 2011 A Robust, Simple Genotyping-by-Sequencing (GBS) Approach for High Diversity Species (L. Orban, Ed.). PLoS ONE 6: e19379.

Endelman, J. B., 2011 Ridge regression and other kernels for genomic selection with R package rrBLUP. Plant Genome 4: 250–255.

Falconer, D. S., T. F. Mackay, and R. Frankham, 1996 Introduction to quantitative genetics (4th edn). Trends Genet. 12: 280.

Fasahat, P., A. Rajabi, J. M. Rad, and J. Derera, 2016 Principles and Utilization of Combining Ability in Plant Breeding. Biom Biostat Int J 4: 00085.

Fehr, W. R., 1987 *Principles of cultivar development. Volume 1. Theory and technique.* Macmillan publishing company.

Fernando, R. L., and D. J. Garrick, 2008 *GenSel-User manual for a portfolio of genomic selection related analyses*.

Friedman, J., T. Hastie, and R. Tibshirani, 2010 Regularization paths for generalized linear models via coordinate descent. J. Stat. Softw. 33: 1.

Fu, D., M. Xiao, A. Hayward, Y. Fu, G. Liu *et al.*, 2014 Utilization of crop heterosis: a review. Euphytica 197: 161–173.

Gianola, D., 2013 Priors in whole-genome regression: the Bayesian alphabet returns. Genetics 194: 573–596.

Gianola, D., G. de los Campos, O. González-Recio, N. Long, H. Okut *et al.*, 2010 Statistical learning methods for genome-based analysis of quantitative traits, pp. 1–6 in *Proceedings of the 9th World Congress on Genetics Applied to Livestock Production*, Gesellschaft für Tierzuchtwissenschaften eV.

Gianola, D., R. L. Fernando, and A. Stella, 2006 Genomic-assisted prediction of genetic value with semiparametric procedures. Genetics 173: 1761–1776.

Gianola, D., and J. B. van Kaam, 2008 Reproducing kernel Hilbert spaces regression methods for genomic assisted prediction of quantitative traits. Genetics 178: 2289–2303.

Gianola, D., H. Okut, K. A. Weigel, and G. J. Rosa, 2011 Predicting complex quantitative traits with Bayesian neural networks: a case study with Jersey cows and wheat. BMC Genet. 12: 87.

Glaubitz, J. C., T. M. Casstevens, F. Lu, J. Harriman, R. J. Elshire *et al.*, 2014 TASSEL-GBS: A High Capacity Genotyping by Sequencing Analysis Pipeline (N. A. Tinker, Ed.). PLoS ONE 9: e90346.

González-Recio, O., G. J. Rosa, and D. Gianola, 2014 Machine learning methods and predictive ability metrics for genome-wide prediction of complex traits. Livest. Sci. 166: 217–231.

Griffing, B., 1956 A generalized treatment of the use of diallel crosses in quantitative inheritance. Heredity 10: 31–50.

Habier, D., R. L. Fernando, and J. C. M. Dekkers, 2007 The impact of genetic relationship information on genome-assisted breeding values. Genetics 177: 2389–2397.

Habier, D., R. L. Fernando, K. Kizilkaya, and D. J. Garrick, 2011 Extension of the Bayesian alphabet for genomic selection. BMC Bioinformatics 12: 186.

Hake, S., and J. Ross-Ibarra, 2015 The natural history of model organisms: genetic, evolutionary and plant breeding insights from the domestication of maize. Elife 4: e05861.

Hallauer, A. R., 1990 Methods used in developing maize inbreds. Maydica 35: 1–16.

Hallauer, A. R., 1977 Relation between inbred and hybrid traits in maize. Crop Sci. 17: 703–706.

Hallauer, A. R., and E. Lopez-Perez, 1979 Comparisons among testers for evaluating lines of corn, in *Proceedings of the... annual corn and sorghum research conference (USA)*,.

Hallauer, A. R., and J. B. Miranda, 1988 Quantitative genetics in plant breeding. Iowa State.

Hallauer, A. R., W. Russell, and K. Lamkey, 1988 Corn breeding. Corn Corn Improv. 18: 463–564.

Hastie, T., R. Tibshirani, and J. Friedman, 2002 The Elements of Statistical Learning: Data Mining, Inference, and Prediction. Biometrics.

Hayes, B. J., P. J. Bowman, A. J. Chamberlain, and M. E. Goddard, 2009 Invited review: Genomic selection in dairy cattle: Progress and challenges. J. Dairy Sci. 92: 433–443.

He, J., X. Zhao, A. Laroche, Z.-X. Lu, H. Liu *et al.*, 2014 Genotyping-by-sequencing (GBS), an ultimate marker-assisted selection (MAS) tool to accelerate plant breeding. Front. Plant Sci. 5:.

Heffner, E. L., A. J. Lorenz, J.-L. Jannink, and M. E. Sorrells, 2010 Plant breeding with genomic selection: gain per unit time and cost. Crop Sci. 50: 1681–1690.

Heffner, E. L., M. E. Sorrells, and J.-L. Jannink, 2009 Genomic selection for crop improvement. Crop Sci. 49: 1–12.

Henderson, C. R., 1984 Applications of linear models in animal breeding. Guelph Univ. Guelph Google Sch.

Heslot, N., H.-P. Yang, M. E. Sorrells, and J.-L. Jannink, 2012 Genomic selection in plant breeding: a comparison of models. Crop Sci. 52: 146–160.

Holland, J. B., W. E. Nyquist, and C. T. Cervantes-Martínez, 2003 Estimating and interpreting heritability for plant breeding: an update. Plant Breed. Rev. 22: 9–112.

Howard, R., A. L. Carriquiry, and W. D. Beavis, 2014 Parametric and nonparametric statistical methods for genomic selection of traits with additive and epistatic genetic architectures. G3 Genes Genomes Genet. 4: 1027–1046.

Isidro, J., J.-L. Jannink, D. Akdemir, J. Poland, N. Heslot *et al.*, 2015 Training set optimization under population structure in genomic selection. TAG Theor. Appl. Genet. Theor. Angew. Genet. 128: 145.

Jacobson, A., L. Lian, S. Zhong, and R. Bernardo, 2014 General combining ability model for genomewide selection in a biparental cross. Crop Sci. 54: 895–905.

Jannink, J.-L., A. J. Lorenz, and H. Iwata, 2010 Genomic selection in plant breeding: from theory to practice. Brief. Funct. Genomics 9: 166–177.

Jenkins, M. T., and A. M. Brunson, 1932 Methods of estimating inbred lines of maize in crossbred combinations.:

Jiang, Y., and J. C. Reif, 2015 Modeling epistasis in genomic selection. Genetics 201: 759–768.

Kadam, D. C., S. M. Potts, M. O. Bohn, A. E. Lipka, and A. J. Lorenz, 2016 Genomic prediction of single crosses in the early stages of a maize hybrid breeding pipeline. G3 Genes Genomes Genet. 6: 3443–3453.

Kohavi, R., 1995 A study of cross-validation and bootstrap for accuracy estimation and model selection, pp. 1137–1145 in *Ijcai*, Stanford, CA.

Lee, E. A., M. J. Ash, and B. Good, 2007 Re-examining the relationship between degree of relatedness, genetic effects, and heterosis in maize. Crop Sci. 47: 629–635.

Legarra, A., C. Robert-Granié, E. Manfredi, and J.-M. Elsen, 2008 Performance of genomic selection in mice. Genetics 180: 611–618.

Li, Z., and M. J. Sillanpää, 2012 Overview of LASSO-related penalized regression methods for quantitative trait mapping and genomic selection. Theor. Appl. Genet. 125: 419–435.

Lian, L., A. Jacobson, S. Zhong, and R. Bernardo, 2014 Genomewide prediction accuracy within 969 maize biparental populations. Crop Sci. 54: 1514–1522.

Lin, Z., B. J. Hayes, and H. D. Daetwyler, 2014 Genomic selection in crops, trees and forages: a review. Crop Pasture Sci. 65: 1177–1191.

Long, N., D. Gianola, G. J. Rosa, and K. A. Weigel, 2011 Application of support vector regression to genome-assisted prediction of quantitative traits. Theor. Appl. Genet. 123: 1065–1074.

Longin, C. F. H., H. F. Utz, J. C. Reif, T. Wegenast, W. Schipprack *et al.*, 2007 Hybrid maize breeding with doubled haploids: III. Efficiency of early testing prior to doubled haploid production in two-stage selection for testcross performance. Theor. Appl. Genet. 115: 519–527.

Lorenz, A. J., S. Chao, F. G. Asoro, E. L. Heffner, T. Hayashi *et al.*, 2011 2 genomic selection in plant breeding: knowledge and prospects. Adv. Agron. 110: 77–123.

Lorenzana, R. E., and R. Bernardo, 2009 Accuracy of genotypic value predictions for marker-based selection in biparental plant populations. Theor. Appl. Genet. 120: 151–161.

Love, H., and B. Wentz, 1914 Correlations between ear characters and yield in corn. Lab. Pap. 1:.

Maenhout, S., B. De Baets, G. Haesaert, and E. Van Bockstaele, 2007 Support vector machine regression for the prediction of maize hybrid performance. Theor. Appl. Genet. 115: 1003–1013.

Mansfield, B. D., and R. H. Mumm, 2014 Survey of plant density tolerance in US maize germplasm. Crop Sci. 54: 157–173.

Massman, J. M., A. Gordillo, R. E. Lorenzana, and R. Bernardo, 2013 Genomewide predictions from maize single-cross data. Theor. Appl. Genet. 126: 13–22.

Massman, J. M., H.-J. G. Jung, and R. Bernardo, 2013 Genomewide selection versus marker-assisted recurrent selection to improve grain yield and stover-quality traits for cellulosic ethanol in maize. Crop Sci. 53: 58–66.

Matsuoka, Y., Y. Vigouroux, M. M. Goodman, J. Sanchez, E. Buckler *et al.*, 2002 A single domestication for maize shown by multilocus microsatellite genotyping. Proc. Natl. Acad. Sci. 99: 6080–6084.

Melchinger, A. E., 1999 Genetic diversity and heterosis. Genet. Exploit. Heterosis Crops 99–118.

Melchinger, A. E., H. H. Geiger, G. Seitz, and G. A. Schmidt, 1987 Optimum prediction of three-way crosses from single crosses in forage maize (Zea mays L.). Theor. Appl. Genet. 74: 339–345.

Meuwissen, T. H. E., B. J. Hayes, and M. E. Goddard, 2001 Prediction of Total Genetic Value Using Genome-Wide Dense Marker Maps. Genetics 157: 1819–1829.

Mikel, M. A., and J. W. Dudley, 2006 Evolution of North American dent corn from public to proprietary germplasm. Crop Sci. 46: 1193–1205.

Morota, G., and D. Gianola, 2014 Kernel-based whole-genome prediction of complex traits: a review. Front. Genet. 5: 363.

Moser, G., B. Tier, R. E. Crump, M. S. Khatkar, and H. W. Raadsma, 2009 A comparison of five methods to predict genomic breeding values of dairy bulls from genome-wide SNP markers. Genet. Sel. Evol. 41: 56.

Neves, H. H., R. Carvalheiro, and S. A. Queiroz, 2012 A comparison of statistical methods for genomic selection in a mice population. BMC Genet. 13: 100.

Ogutu, J. O., T. Schulz-Streeck, and H.-P. Piepho, 2012 Genomic selection using regularized linear regression models: ridge regression, lasso, elastic net and their extensions, pp. S10 in *BMC proceedings*, BioMed Central.

Patterson, H. D., and E. R. Williams, 1976 A new class of resolvable incomplete block designs. Biometrika 63: 83–92.

Pérez, P., and G. de Los Campos, 2014 Genome-wide regression & prediction with the BGLR statistical package. Genetics genetics–114.

Pérez-Rodriguez, P., and D. Gianola, 2013 *brnn: brnn (Bayesian regularization for feed-forward neural networks). R package version 0.3*.

Pérez-Rodríguez, P., D. Gianola, J. M. González-Camacho, J. Crossa, Y. Manès *et al.*, 2012 Comparison between linear and non-parametric regression models for genome-enabled prediction in wheat. G3 Genes Genomes Genet. 2: 1595–1605.

Pérez-Rodríguez, P., D. Gianola, K. A. Weigel, G. J. M. Rosa, and J. Crossa, 2013 An R package for fitting Bayesian regularized neural networks with applications in animal breeding. J. Anim. Sci. 91: 3522–3531.

Piepho, H.-P., 2009 Ridge regression and extensions for genomewide selection in maize. Crop Sci. 49: 1165–1176.

Poland, J., J. Endelman, J. Dawson, J. Rutkoski, S. Wu *et al.*, 2012 Genomic selection in wheat breeding using genotyping-by-sequencing. Plant Genome 5: 103–113.

Ranum, P., J. P. Peña-Rosas, and M. N. Garcia-Casal, 2014 Global maize production, utilization, and consumption. Ann. N. Y. Acad. Sci. 1312: 105–112.

Reif, J. C., F.-M. Gumpert, S. Fischer, and A. E. Melchinger, 2007 Impact of interpopulation divergence on additive and dominance variance in hybrid populations. Genetics 176: 1931–1934.

Reif, J. C., A. R. Hallauer, and A. E. Melchinger, 2005 Heterosis and heterotic patterns in maize. Maydica 50: 215.

Riedelsheimer, C., A. Czedik-Eysenberg, C. Grieder, J. Lisec, F. Technow *et al.*, 2012 Genomic and metabolic prediction of complex heterotic traits in hybrid maize. Nat. Genet. 44: 217–220.

Riedelsheimer, C., J. B. Endelman, M. Stange, M. E. Sorrells, J.-L. Jannink *et al.*, 2013 Genomic predictability of interconnected biparental maize populations. Genetics 194: 493–503.

Rincent, R., D. Laloë, S. Nicolas, T. Altmann, D. Brunel *et al.*, 2012 Maximizing the reliability of genomic selection by optimizing the calibration set of reference individuals: comparison of methods in two diverse groups of maize inbreds (Zea mays L.). Genetics 192: 715–728.

Rober, F. K., G. A. Gordillo, and H. H. Geiger, 2005 In vivo haploid induction in maize-performance of new inducers and significance of doubled haploid lines in hybrid breeding. Maydica 50: 275.

Schrag, T. A., M. Frish, B. S. Dhillon, and A. E. Melchinger, 2009 Marker-based prediction of hybrid performance in maize single-crosses involving doubled haploids. Maydica 54: 353.

Schrag, T. A., H. P. Maurer, A. E. Melchinger, H.-P. Piepho, J. Peleman *et al.*, 2007 Prediction of single-cross hybrid performance in maize using haplotype blocks associated with QTL for grain yield. Theor. Appl. Genet. 114: 1345–1355.

Schrag, T. A., A. E. Melchinger, A. P. Sørensen, and M. Frisch, 2006 Prediction of single-cross hybrid performance for grain yield and grain dry matter content in maize using AFLP markers associated with QTL. Theor. Appl. Genet. 113: 1037–1047.

Schulz-Streeck, T., J. O. Ogutu, Z. Karaman, C. Knaak, and H. P. Piepho, 2012 Genomic selection using multiple populations. Crop Sci. 52: 2453–2461.

Shiferaw, B., B. M. Prasanna, J. Hellin, and M. Bänziger, 2011 Crops that feed the world 6. Past successes and future challenges to the role played by maize in global food security. Food Secur. 3: 307–327.

Shull, G. H., 1909 A pure-line method in corn breeding. J. Hered. 51–58.

Shull, G. H., 1952 Beginnings of the heterosis concept. Begin. Heterosis Concept.

Shull, G. H., 1908 The composition of a field of maize.

Smith, C. W., 2004 *Corn: origin, history, technology, and production*. John Wiley & Sons.

Smith, O. S., 1986 Covariance between line per se and testcross performance. Crop Sci. 26: 540–543.

Solberg, T. R., A. K. Sonesson, J. A. Woolliams, and others, 2008 Genomic selection using different marker types and densities. J. Anim. Sci. 86: 2447–2454.

Sprague, G. F., and L. A. Tatum, 1942 General vs. specific combining ability in single crosses of corn. Agron. J. 34: 923–932.

SPRAGUEI, O., and L. A. Tatum, 1942 General versus specific combining ability in single crosses of com. J. Am. Soc. Agron. Wash. 34: 923–932.

Strable, J., and M. J. Scanlon, 2009 Maize (Zea mays): a model organism for basic and applied research in plant biology. Cold Spring Harb. Protoc. 2009: pdb–emo132.

Su, G., O. F. Christensen, T. Ostersen, M. Henryon, and M. S. Lund, 2012 Estimating additive and non-additive genetic variances and predicting genetic merits using genome-wide dense single nucleotide polymorphism markers. PloS One 7: e45293.

Sun, C., P. M. VanRaden, J. B. Cole, and J. R. O'Connell, 2014 Improvement of prediction ability for genomic selection of dairy cattle by including dominance effects. PloS One 9: e103934.

Technow, F., A. Bürger, and A. E. Melchinger, 2013 Genomic prediction of northern corn leaf blight resistance in maize with combined or separated training sets for heterotic groups. G3 Genes Genomes Genet. 3: 197–203.

Technow, F., C. Riedelsheimer, T. A. Schrag, and A. E. Melchinger, 2012 Genomic prediction of hybrid performance in maize with models incorporating dominance and population specific marker effects. Theor. Appl. Genet. 1–14.

Technow, F., T. A. Schrag, W. Schipprack, E. Bauer, H. Simianer *et al.*, 2014 Genome Properties and Prospects of Genomic Prediction of Hybrid Performance in a Breeding Program of Maize. Genetics 197: 1343–1355.

Thomson, M. J., 2014 High-throughput SNP genotyping to accelerate crop improvement. Plant Breed. Biotechnol. 2: 195–212.

VanRaden, P. M., 2008 Efficient methods to compute genomic predictions. J. Dairy Sci. 91: 4414–4423.

Viana, J. M. S., H.-P. Piepho, and others, 2016 Quantitative genetics theory for genomic selection and efficiency of breeding value prediction in open-pollinated populations. Sci. Agric. 73: 243–251.

Vivek, B. S., G. K. Krishna, V. Vengadessan, R. Babu, P. H. Zaidi *et al.*, 2017 Use of genomic estimated breeding values results in rapid genetic gains for drought tolerance in maize. Plant Genome 10:.

Vuylsteke, M., M. Kuiper, and P. Stam, 2000 Chromosomal regions involved in hybrid performance and heterosis: their AFLP®-based identification and practical use in prediction models. Heredity 85: 208–218.

Wegenast, T., C. F. H. Longin, H. F. Utz, A. E. Melchinger, H. P. Maurer *et al.*, 2008 Hybrid maize breeding with doubled haploids. IV. Number versus size of crosses and importance of parental selection in two-stage selection for testcross performance. Theor. Appl. Genet. 117: 251–260.

Windhausen, V. S., G. N. Atlin, J. M. Hickey, J. Crossa, J.-L. Jannink *et al.*, 2012 Effectiveness of Genomic Prediction of Maize Hybrid Performance in Different

Breeding Populations and Environments. G3amp58 GenesGenomesGenetics 2: 1427–1436.

Wu, X.-L., T. M. Beissinger, S. Bauck, B. Woodward, G. J. Rosa *et al.*, 2011 A primer on high-throughput computing for genomic selection. Front. Genet. 2:.

Xu, S., D. Zhu, and Q. Zhang, 2014 Predicting hybrid performance in rice using genomic best linear unbiased prediction. Proc. Natl. Acad. Sci. 111: 12456–12461.

Zeileis, A., K. Hornik, A. Smola, and A. Karatzoglou, 2004 kernlab-an S4 package for kernel methods in R. J. Stat. Softw. 11: 1–20.

Zenke-Philippi, C., A. Thiemann, F. Seifert, T. Schrag, A. E. Melchinger *et al.*, 2016 Prediction of hybrid performance in maize with a ridge regression model employed to DNA markers and mRNA transcription profiles. BMC Genomics 17: 262.

Zhang, X., P. Pérez-Rodríguez, J. Burgueño, M. Olsen, E. Buckler *et al.*, 2017 Rapid Cycling Genomic Selection in a Multi-parental Tropical Maize Population. G3 Genes Genomes Genet. g3–117.

Zhao, Y., M. Gowda, W. Liu, T. Würschum, H. P. Maurer *et al.*, 2012 Accuracy of genomic selection in European maize elite breeding populations. Theor. Appl. Genet. 124: 769–776.

Zhao, Y., Z. Li, G. Liu, Y. Jiang, H. P. Maurer *et al.*, 2015 Genome-based establishment of a high-yielding heterotic pattern for hybrid wheat breeding. Proc. Natl. Acad. Sci. 201514547.

Zhao, Y., J. Zeng, R. Fernando, and J. C. Reif, 2013 Genomic prediction of hybrid wheat performance. Crop Sci. 53: 802–810.

Zhong, S., J. C. Dekkers, R. L. Fernando, and J.-L. Jannink, 2009 Factors affecting accuracy from genomic selection in populations derived from multiple inbred lines: a barley case study. Genetics 182: 355–364.

Zou, H., and T. Hastie, 2005 Regularization and variable selection via the elastic net. J. R. Stat. Soc. Ser. B Stat. Methodol. 67: 301–320.